

# ADAPTIVE PARALLEL TEMPERING ALGORITHM

BLAŻEJ MIASOJEDOW, ERIC MOULINES AND MATTI VIHOLA

**ABSTRACT.** Parallel tempering is a generic Markov chain Monte Carlo sampling method which allows good mixing with multimodal target distributions, where conventional Metropolis-Hastings algorithms often fail. The mixing properties of the sampler depend strongly on the choice of tuning parameters, such as the temperature schedule and the proposal distribution used for local exploration. We propose an adaptive algorithm which tunes both the temperature schedule and the parameters of the random-walk Metropolis kernel automatically. We prove the convergence of the adaptation and a strong law of large numbers for the algorithm. We illustrate the performance of our method with examples. Our empirical findings indicate that the algorithm can cope well with different kind of scenarios without prior tuning.

## 1. INTRODUCTION

Markov chain Monte Carlo (MCMC) is a generic method to approximate an integral of the form

$$I \stackrel{\text{def}}{=} \int_{\mathbb{R}^d} f(x) \pi(x) dx,$$

where  $\pi$  is a probability density function, which can be evaluated pointwise up to a normalising constant. Such an integral occurs frequently when computing Bayesian posterior expectations [e.g., 11, 24].

The random-walk Metropolis algorithm [21] works often well, provided the target density  $\pi$  is, roughly speaking, sufficiently close to unimodal. The efficiency of the Metropolis algorithm can be optimised by a suitable choice of the proposal distribution. The proposal distribution can be chosen automatically by several adaptive MCMC algorithms; see [2, 4, 12, 26] and references therein.

When  $\pi$  has multiple well-separated modes, the random-walk based methods tend to get stuck to a single mode for long periods of time, which can lead to false convergence and severely erroneous results. Using a tailored Metropolis-Hastings algorithm can help, but in many cases finding a good proposal distribution is difficult. Tempering of  $\pi$ , that is, considering auxiliary distributions with density proportional to  $\pi^\beta$  with  $\beta \in (0, 1)$  often provides better mixing within modes [13, 20, 30, 34]. We focus here particularly on the parallel tempering algorithm, which is also known as the replica exchange Monte Carlo and the Metropolis coupled Markov chain Monte Carlo.

The tempering approach is particularly tempting in such settings where  $\pi$  admits a physical interpretation, and there is good intuition how to choose the temperature schedule for the algorithm. In general, choosing the temperature schedule is a non-trivial task, but there are generic guidelines for temperature selection, based on both empirical findings and theoretical analysis. First rule of thumb suggests that the temperature progression should be (approximately) geometric; see, e.g. [16]. Kone and Kofke linked also the mean acceptance rate of the swaps [17]; this has been further analysed by Atchadé, Roberts and Rosenthal [5]; see also [27].

Our temperature adaptation is based on the latter; we try to optimise the mean acceptance rate of the swaps between the chains in adjacent temperatures. Our scheme has similarities with that proposed by Atchadé, Roberts and Rosenthal [5]. The key difference in our method is that we propose to adapt continuously during the simulation. We show that the temperature adaptation converges, and that the point of convergence is unique with mild and natural conditions on the target distribution.

The local exploration in our approach relies on the random walk Metropolis algorithm. The proposal distribution, or more precisely, the scale/shape parameter of the proposal distribution, can be adapted using several existing techniques like the covariance estimator [12] augmented with an adaptive scaling pursuing a given mean acceptance rate [2, 3, 4, 26] which is motivated by certain asymptotic results [25, 28]. It is also possible to use a robust shape estimate which enforces a given mean acceptance rate [33].

We start by describing the proposed algorithm in Section 2. Theoretical results on the convergence of the adaptation and the ergodic averages are given next in Section 3. In Section 4, we illustrate the performance of the algorithm with examples. The proofs of the theoretical results are postponed to Section 5.

## 2. ALGORITHM

**2.1. Parallel tempering algorithm.** The parallel tempering algorithm defines a Markov chain over the product space  $\mathbf{X}^L$ , where  $\mathbf{X} \subset \mathbb{R}^d$

$$(1) \quad \mathbf{X}_k = (X_k^{(1)}, \dots, X_k^{(L)}) = (X_k^{(1:L)}).$$

Each of the chains  $X_k^{(\ell)}$  targets a ‘tempered’ version of the target distribution  $\pi$ . Denote by  $\boldsymbol{\beta} = (\beta^{(1:L)})$  the inverse temperature, which are such that  $1 = \beta^{(1)} > \beta^{(2)} > \dots > \beta^{(L)} > 0$ . and by  $Z(\beta)$  the normalising constant

$$(2) \quad Z(\beta) \stackrel{\text{def}}{=} \int \pi^\beta(x) dx,$$

which is assumed to be finite. The parallel tempering algorithm is constructed so that the Markov chain  $\{\mathbf{X}_k\}_{k \geq 0}$  is reversible with respect to the product density

$$(3) \quad \boldsymbol{\pi}_\beta(x^{(1)}, \dots, x^{(L)}) \stackrel{\text{def}}{=} \frac{\pi^{\beta^{(1)}}(x^{(1)})}{Z(\beta^{(1)})} \times \dots \times \frac{\pi^{\beta^{(L)}}(x^{(L)})}{Z(\beta^{(L)})},$$

over the product space  $(\mathbf{X}^L, \mathcal{X}^{\otimes L})$ .

Each time-step may be decomposed into two successive moves: the swap move and the propagation (or update) move; for the latter, we consider only random-walk Metropolis moves.

We use the following notation to distinguish the state of the algorithm after the swap step (denoted  $\bar{\mathbf{X}}_n$ ) and after the random walk step, or equivalently after a complete step (denoted  $\mathbf{X}_n$ ). The state is then updated according to

$$(4) \quad \mathbf{X}_{n-1} \xrightarrow{\mathbf{S}_\beta} \bar{\mathbf{X}}_{n-1} \xrightarrow{\mathbf{M}_{(\boldsymbol{\Sigma}, \beta)}} \mathbf{X}_n,$$

where the two kernels  $\mathbf{M}_{(\boldsymbol{\Sigma}, \beta)}$  and  $\mathbf{S}_\beta$  are respectively defined as

- $\mathbf{M}_{(\boldsymbol{\Sigma}, \beta)}$  denotes the tensor product kernel on the product space  $\mathbf{X}^L$

$$(5) \quad \mathbf{M}_{(\boldsymbol{\Sigma}, \beta)}(x^{(1:L)}; A_1 \times \dots \times A_L) = \prod_{\ell=1}^L M_{(\boldsymbol{\Sigma}^{(\ell)}, \beta^{(\ell)})}(x^{(\ell)}, A_\ell)$$

where each  $M_{(\boldsymbol{\Sigma}^{(\ell)}, \beta^{(\ell)})}$  is a random-walk Metropolis kernel targeting  $\pi^{\beta^{(\ell)}}$  with increment distribution  $q_{\boldsymbol{\Sigma}^{(\ell)}}$ , where  $q_\Sigma$  is the density of a multivariate Gaussian with zero mean and covariance  $\Sigma$ ,

$$(6) \quad M_{(\boldsymbol{\Sigma}, \beta)}(x, A) \stackrel{\text{def}}{=} \int_A \alpha_\beta(x, y) q_\Sigma(y - x) dy + \delta_x(A) \int (1 - \alpha_\beta(x, y)) q_\Sigma(y - x) dy,$$

where

$$(7) \quad \alpha_\beta(x, y) \stackrel{\text{def}}{=} 1 \wedge \frac{\pi^\beta(y)}{\pi^\beta(x)}, \quad \text{for all } (x, y) \in \mathbf{X} \times \mathbf{X}.$$

In practical terms,  $\mathbf{M}_{(\boldsymbol{\Sigma}, \beta)}$  means that one applies a random-walk Metropolis step separately for each of the chains  $\bar{X}_{n-1}^{(\ell)}$ .

- $\mathbf{S}_\beta$  denotes the Markov kernel of the swap steps, targeting the product distribution  $\boldsymbol{\pi}_\beta \propto \pi^{\beta(1)} \otimes \cdots \otimes \pi^{\beta(L)}$ ,

$$(8) \quad \mathbf{S}_\beta(x^{(1:L)}; A) = \frac{1}{L-1} \sum_{j=1}^{L-1} \varpi_\beta^{(j)}(x^{(j)}, x^{(j+1)}) J^{(j)}(x^{(j)}, x^{(j+1)}; A) \\ + \frac{1}{L-1} \sum_{j=1}^{L-1} (1 - \varpi_\beta^{(j)}(x^{(j)}, x^{(j+1)})) \delta_{x^{(1:L)}}(A),$$

where  $\varpi_\beta^{(j)}$  is the probability of accepting a swap between levels  $j$  and  $j+1$ , which is given by

$$(9) \quad \varpi_\beta^{(j)}(x^{(j)}, x^{(j+1)}) \stackrel{\text{def}}{=} 1 \wedge \left( \frac{\pi(x^{(j+1)})}{\pi(x^{(j)})} \right)^{\beta^{(j)} - \beta^{(j+1)}},$$

and

$$(10) \quad J^{(j)}(x^{(j)}, x^{(j+1)}; A) \\ \stackrel{\text{def}}{=} \int \cdots \int_A \delta_{x^{(j)}}(dy^{(j+1)}) \delta_{x^{(j+1)}}(dy^{(j)}) \prod_{i \in \{1, \dots, L\} \setminus \{j, j+1\}} \delta_{x^{(i)}}(dy^{(i)}).$$

The above defined swap step means choosing a random index  $\ell \in \{1, \dots, L-1\}$  uniformly, and then proposing to swap the adjacent states,  $\bar{X}_{n-1}^{(\ell+1)} = X_{n-1}^{(\ell)}$  and  $\bar{X}_{n-1}^{(\ell)} = X_{n-1}^{(\ell+1)}$ . and accepting this swap with probability given in (9).

**2.2. Adaptive parallel tempering algorithm.** In the adaptive version of the parallel tempering algorithm, the temperature parameters are continuously updated along the run of the algorithm. We denote the sequence of inverse temperatures

$$(11) \quad \{\boldsymbol{\beta}_n\}_{n \geq 0} = \{\beta_n^{(1:L)}\}_{n \geq 0}$$

which are parameterised by the vector-valued process

$$(12) \quad \{\boldsymbol{\rho}_n\}_{n \geq 0} \stackrel{\text{def}}{=} \{\rho_n^{(1:L-1)}\}_{n \geq 0},$$

through  $\beta_n^{(1)} \stackrel{\text{def}}{=} 1$  and  $\beta_n^{(\ell)} = \beta^{(\ell)}(\rho_n^{(1:\ell-1)})$  for  $\ell \in \{2, \dots, L\}$  with

$$(13) \quad \beta^{(\ell+1)}(\rho^{(1:\ell)}) \stackrel{\text{def}}{=} \beta^{(\ell)}(\rho^{(1:\ell-1)}) \exp(-\exp(\rho^{(\ell)})).$$

Because the inverse temperatures are adapted at each iteration, the target distribution of the chain changes from step to step as well. Our adaptation of the temperatures is performed using the following stochastic approximation procedure

$$(14) \quad \rho_n^{(\ell)} = \Pi_\rho \left( \rho_{n-1}^{(\ell)} + \gamma_{n,1} H^{(\ell)}(\rho_{n-1}^{(1:\ell)}, \mathbf{X}_n) \right), \quad 1 \leq \ell \leq L-1,$$

where  $(\rho_n^{(1:L-1)})$  is defined in (13),  $\Pi_\rho$  is the projection onto the constraint set  $[\underline{\rho}, \overline{\rho}]$ , which will be discussed further in Section 2.4. Moreover,

$$(15) \quad H^{(\ell)}(\rho^{(1:\ell)}, \mathbf{x}) = 1 \wedge \left( \frac{\pi(x^{(\ell+1)})}{\pi(x^{(\ell)})} \right)^{\Delta\beta^{(\ell)}(\rho^{(1:\ell)})} - \alpha^*,$$

$$(16) \quad \Delta\beta^{(\ell)}(\rho^{(1:\ell)}) = \beta^{(\ell)}(\rho^{(1:\ell-1)}) - \beta^{(\ell+1)}(\rho^{(1:\ell)}).$$

We will show in Section 3 that the algorithm is designed in such a way that the inverse temperatures converges to a value for which the mean probability of accepting a swap move between any adjacent-temperature chains is constant and is equal to  $\alpha^*$ .

We will also adapt the random-walk proposal distribution at each level. We describe below another possible algorithm for performing such a task. In the theoretical part, for simplicity, we will consider only on with the seminal adaptive Metropolis algorithm [12] augmented with scaling adaptation [e.g. 2, 3, 26]. In this algorithm, we estimate the covariance matrix of the target distribution at each temperature and rescale it to control the acceptance ratio at each level in stationarity.

Define by  $\mathcal{M}_+(d)$  the set of  $d \times d$  positive definite matrices. For  $A \in \mathcal{M}_+(d)$ , we denote by  $\varrho_{\min}(A)$  and  $\varrho_{\max}(A)$  the smallest and the largest eigenvalues of  $A$ , respectively. For  $\varepsilon \in (0, 1]$ , define by  $\mathcal{M}_+(d, \varepsilon) \subset \mathcal{M}_+(d)$  the convex subset

$$(17) \quad \mathcal{M}_+(d, \varepsilon) \stackrel{\text{def}}{=} \{ \Sigma \in \mathcal{M}_+(d) : \varepsilon \leq \varrho_{\min}(\Sigma) \leq \varrho_{\max}(\Sigma) \leq \varepsilon^{-1} \}.$$

The set  $\mathcal{M}_+(d, \varepsilon)$  is a compact subset of the open cone of positive definite matrices.

We denote by  $\Gamma_n^{(\ell)}$  the current estimate of the covariance at level  $\ell$ , which is updated as follows

$$(18) \quad \Gamma_n^{(\ell)} = \Pi_\Gamma \left[ (1 - \gamma_{n,2}) \Gamma_{n-1}^{(\ell)} + \gamma_{n,2} (X_n^{(\ell)} - \mu_{n-1}^{(\ell)}) t(X_n^{(\ell)} - \mu_{n-1}^{(\ell)}) \right],$$

$$(19) \quad \mu_n^{(\ell)} = (1 - \gamma_{n,2}) \mu_{n-1}^{(\ell)} + \gamma_{n,2} X_n^{(\ell)},$$

where  $t(\cdot)$  is the matrix transpose and  $\Pi_\Gamma$  is the projection on to the set  $\mathcal{M}_+(d, \varepsilon)$ ; see Section 2.4. The scaling parameters is updated so that the acceptance rate in stationarity converges to the target  $\alpha^*$ ,

$$(20) \quad T_n^{(\ell)} = \Pi_T \left( T_{n-1}^{(\ell)} + \gamma_{n,3} \left[ \left( 1 \wedge \frac{\pi^{\beta_{n-1}^{(\ell)}}(Y_n^{(\ell)})}{\pi^{\beta_{n-1}^{(\ell)}}(\bar{X}_{n-1}^{(\ell)})} \right) - \alpha^* \right] \right),$$

where  $\Pi_T$  is the projection onto  $[\underline{T}, \overline{T}]$ ; see Section 2.4 and  $Y_n^{(\ell)}$  is the proposal at level  $\ell$ , assumed to be conditionally independent from the past draws and distributed according to a Gaussian with mean  $\bar{X}_{n-1}^{(\ell)}$  and covariance matrix  $\Sigma_{n-1}^{(\ell)}$  which is given by

$$(21) \quad \Sigma_n^{(\ell)} = \exp(T_n^{(\ell)}) \Gamma_n^{(\ell)}.$$

In the sequel we denote by  $\mathbf{Y}_n$  the vector of proposed moves at time step  $n$ ,

$$(22) \quad \{\mathbf{Y}_n\}_{n \geq 0} = \{Y_n^{(1:L)}\}_{n \geq 0}.$$

**2.3. Alternate random-walk adaptation.** In order to reduce the number of parameters in the adaptation especially in higher dimensions, we propose to use a common covariance for all the temperatures, but still employ separate scaling. More specifically,

$$(23) \quad \Gamma_n = (1 - \gamma_{n,2})\Gamma_{n-1} + \frac{\gamma_{n,2}}{L} \sum_{\ell=1}^L (X_n^{(\ell)} - \mu_{n-1})t(X_n^{(\ell)} - \mu_{n-1}),$$

$$(24) \quad \mu_n = (1 - \gamma_{n,2})\mu_{n-1} + \frac{\gamma_{n,2}}{L} \sum_{\ell=1}^L X_n^{(\ell)},$$

and set  $\Sigma_n^{(\ell)} = \exp(T_n^{(\ell)})\Gamma_n$ .

Another possible implementation of the random-walk adaption, robust adaptive Metropolis (RAM) [33], is defined by a single dynamic adjusting the covariance parameter and attaining a given acceptance rate. Specifically, one recursively finds a lower-diagonal matrix  $\Gamma_n^{(\ell)} \in \mathbb{R}^{d \times d}$  with positive diagonal satisfying

$$(25) \quad \Gamma_n^{(\ell)} t(\Gamma_n^{(\ell)}) = \Gamma_{n-1}^{(\ell)} [I + \gamma_{n,2}(\alpha_n - \alpha^*)u(Z_n^{(\ell)})t(u(Z_n^{(\ell)}))] t(\Gamma_{n-1}^{(\ell)}),$$

where  $Z_n^{(\ell)} \stackrel{\text{def}}{=} Y_n^{(\ell)} - \bar{X}_{n-1}^{(\ell)}$  and  $u(x) := \mathbb{I}\{x \neq 0\} x/|x|$ , and let  $\Sigma_n^{(\ell)} = \Gamma_n^{(\ell)} t(\Gamma_n^{(\ell)})$ .

The potential benefit of using this estimate instead of (18)–(20) is that RAM finds, loosely speaking, a ‘local’ shape of the target distribution, which is often in practice close to a convex combination of the shapes of individual modes. In some situations, this proposal shape might allow better local exploration than the global covariance shape.

**2.4. Implementation details.** In the experiments, we use the desired acceptance rate  $\alpha^* = 0.234$  suggested by theoretical results for the swap kernel [5, 17] and for the random-walk Metropolis kernel [25, 28]. We employ the step size sequences  $\gamma_{n,i} = c_i(n+1)^{-\xi_i}$  with constants  $c_1, c_3 \in (0, \infty)$  and  $c_2 \in (0, 1]$  and  $\xi_1, \xi_2, \xi_3 \in (1/2, 1)$ . This is a common choice in the stochastic approximation literature.

The projections  $\Pi_\rho$ ,  $\Pi_\Gamma$  and  $\Pi_T$  in (14), (18) and (20), respectively, are used to enforce the stability of the adaptation process in order to simplify theoretical analysis of the algorithm. We have not observed instability empirically, and believe that the algorithm would be stable without projections; in fact, for the random-walk adaptation, there exist some stability results [29, 31, 32]. Therefore, we recommend setting the limits in the constraint sets as large as possible, within the limits of numerical accuracy.

It is possible to employ other strategies for proposing swaps of the tempered states. Specifically, it is possible to try more than one swap at each iteration, even go through all the temperatures, without changing the invariant distribution of the chain. We made some preliminary tests with other strategies, but the results were not promising, so we decided to keep the common approach of a single randomly chosen swap.

In the temperature adaptation, it is also possible to enforce the geometric progression, and only adapt one parameter. More specifically, one can use  $\rho_n^{(\ell)} := \rho_n$  for all  $\ell \in \{1, \dots, L-1\}$  and perform the adaptation (14) to update  $\rho_{n-1} \rightarrow \rho_n$ . This strategy might induce more stable behaviour of the temperature parameter especially when the number of levels is high. On the other hand, it can be dangerous because the asymptotic acceptance probability across certain temperature levels can get low, inducing poor mixing.

We consider only Gaussian proposal distributions in the random-walk Metropolis kernels. It is possible to employ also other proposals; in fact our theoretical results extend directly for example to the multivariate Student proposal distributions.

We note that the adaptive parallel tempering algorithm can be used also in a block-wise manner, or in Metropolis-within-Gibbs framework. More precisely, the adaptive random-walk chains can be run as Metropolis-within-Gibbs, and the state swapping can be done in the global level. This approach scales better with respect to the dimension in many situations. Particularly, when the model is hierarchical, the structure of the model can allow significant computational savings. Finally, it is straightforward to extend the adaptive parallel tempering algorithm described above to general measure spaces. For the sake of exposition, we present the algorithm only in  $\mathbb{R}^d$ .

### 3. THEORETICAL RESULTS

**3.1. Formal definitions and assumptions.** Denote by  $\{Y_n\}$  the proposals of the random-walk Metropolis step. We define the following filtration

$$(26) \quad \mathcal{F}_n = \sigma \{ \mathbf{X}_0, (\mathbf{X}_k, \bar{\mathbf{X}}_{k-1}, \mathbf{Y}_{k-1}), k = 1, \dots, n, \} .$$

By construction, the covariance matrix  $\Sigma_n \stackrel{\text{def}}{=} (\Sigma_n^{(1:L)})$  and the parameters  $\beta_n \stackrel{\text{def}}{=} \beta_n^{(1:L)}$  are adapted to the filtration  $\mathcal{F}_n$ . With these notations and assumptions, for any time step  $n \in \mathbb{N}$ ,

$$\mathbb{P} [\mathbf{X}_{n+1} \in \cdot | \mathcal{F}_n] = \int \mathbf{S}_{\beta_n}(\mathbf{X}_n, dz) \mathbf{M}_{(\Sigma_n, \beta_n)}(z, \cdot) = \mathbf{S}_{\beta_n} \mathbf{M}_{(\Sigma_n, \beta_n)}(\mathbf{X}_n, \cdot)$$

Therefore, denoting  $\mathbf{P}_{(\Sigma_n, \beta_n)} \stackrel{\text{def}}{=} \mathbf{S}_{\beta_n} \mathbf{M}_{(\Sigma_n, \beta_n)}$ , we get

$$(27) \quad \mathbb{E} [f(\mathbf{X}_{n+1}) | \mathcal{F}_n] = \mathbf{P}_{(\Sigma_n, \beta_n)} f(\mathbf{X}_n) ,$$

for all  $n \in \mathbb{N}$  and all bounded measurable functions  $f : \mathbf{X}^L \rightarrow \mathbb{R}$ .

We will consider the following assumption on the target distribution, which ensures a geometric ergodicity of a random walk Metropolis chain [1, 15]. Below,  $|\cdot|$  applied to a vector (or a matrix) stands for the Euclidean norm.

(A1) The density  $\pi$  is bounded, bounded away from zero on compact sets, differentiable, such that

$$(28) \quad \lim_{r \rightarrow \infty} \sup_{|x| \geq r} \frac{x}{|x|} \cdot \nabla \log \pi(x) = -\infty$$

$$(29) \quad \lim_{r \rightarrow \infty} \sup_{|x| \geq r} \frac{x}{|x|} \cdot \frac{\nabla \pi(x)}{|\nabla \pi(x)|} < 0.$$

In words, (A1) only requires that the target distribution is sufficiently regular, and the tails decay at a rate faster than exponential. We remark that the tempering approach is only well-defined when  $\pi^\beta$  are integrable with exponents of interest  $\beta > 0$ —this is the case always with (A1).

**3.2. Geometric ergodicity and continuity of parallel tempering kernels.** We first state and prove that the parallel tempering algorithm is geometrically ergodic under (A1). This result might be of independent interest, because geometric ergodicity is well known to imply central limit theorems.

We show that, under mild conditions, this kernel is  $\phi$ -irreducible, strongly aperiodic, and  $V$ -uniformly ergodic, where the function  $V$  is the sum of an appropriately chosen negative power of the target density. Specifically, for  $\beta \in \mathbb{R}_+$ , consider the following drift function

$$(30) \quad \mathbf{V}_\beta(x^{(1:L)}) \stackrel{\text{def}}{=} \sum_{\ell=1}^L V_\beta(x^{(\ell)}),$$

where for  $x \in \mathbf{X}$ ,

$$(31) \quad V_\beta(x) = (\pi(x) / \|\pi\|_\infty)^{-\beta/2}.$$

For  $\beta_0 > 0$ , define the set

$$(32) \quad \mathcal{K}_{\beta_0} \stackrel{\text{def}}{=} \left\{ \beta^{(1:L)} \in (0, 1]^L, \beta_0 \leq \beta^{(L)} \leq \dots \leq \beta^{(1)} \right\}.$$

We denote the  $V$ -variation of a signed measure  $\mu$  as  $\|\mu\|_V := \sup_{f: |f| \leq V} \mu(f)$ , where the supremum is taken over all measurable functions  $f$ . The  $V$ -norm of a function is defined as  $\|f\|_V \stackrel{\text{def}}{=} \sup_x |f(x)|/V(x)$ .

**Theorem 1.** *Assume (A1). Let  $\epsilon > 0$  and  $\beta_0 > 0$ . Then there exists  $C_{\epsilon, \beta_0} < \infty$  and  $\varrho_{\epsilon, \beta_0} < 1$  such that, for all  $\mathbf{x} \in \mathbf{X}^L$ ,  $\Sigma \in \mathcal{M}_+(d, \epsilon)$  and  $\beta \in \mathcal{K}_{\beta_0}$ ,*

$$(33) \quad \left\| \mathbf{P}_{(\Sigma, \beta)}^n(\mathbf{x}, \cdot) - \pi_\beta \right\|_{\mathbf{V}_{\beta_0}} \leq C_{\epsilon, \beta_0} \varrho_{\epsilon, \beta_0}^n \mathbf{V}_{\beta_0}(x).$$



Geometric ergodicity in turn implies the existence of a solution of the Poisson equation, and also provides bounds on the growth of this solution [22, Chapter 17]

**Corollary 2.** *Assume (A1). Let  $\epsilon > 0$  and  $\beta_0 > 0$ . For any measurable function  $f$  with  $\|f\|_{\mathbf{V}_{\beta_0}^\alpha} < \infty$  for some  $\alpha \in (0, 1]$  there exists a unique (up to an additive constant) solution of the Poisson equation*

$$(34) \quad g - \mathbf{P}_{(\Sigma, \beta)} g = f - \pi_\beta(f).$$

*This solution is denoted  $\hat{f}_{(\Sigma, \beta)}$ . In addition, there exists a constant  $D_{\epsilon, \beta_0} < \infty$  such that*

$$(35) \quad \|\hat{f}_{(\Sigma, \beta)}\|_{\mathbf{V}_{\beta_0}^\alpha} \leq D_{\epsilon, \beta_0} \|f\|_{\mathbf{V}_{\beta_0}^\alpha}.$$

We will next establish that the parallel tempering kernel is locally Lipschitz continuous. For any  $\beta > 0$ , denote by  $D_{\mathbf{V}_\beta}[(\Sigma, \beta), (\Sigma', \beta')]$  the  $\mathbf{V}_\beta$ -variation of the kernels  $\mathbf{P}_{(\Sigma, \beta)}$  and  $\mathbf{P}_{(\Sigma', \beta')}$ ,

$$(36) \quad D_{\mathbf{V}_\beta}[(\Sigma, \beta), (\Sigma', \beta')] \stackrel{\text{def}}{=} \sup_{\mathbf{x} \in \mathbf{X}^L} \frac{\|\mathbf{P}_{(\Sigma, \beta)}(\mathbf{x}, \cdot) - \mathbf{P}_{(\Sigma', \beta')}(\mathbf{x}, \cdot)\|_{\mathbf{V}_\beta}}{\mathbf{V}_\beta(\mathbf{x})}.$$

For  $\beta_0 \in (0, 1)$  and  $\eta > 0$ , define the set

$$(37) \quad \mathcal{K}_{\beta_0, \eta} \stackrel{\text{def}}{=} \{\beta_0 \leq \beta^{(L)} \leq \dots \leq \beta^{(1)} \leq 1, \beta^{(k)} - \beta^{(k+1)} \geq \eta\}.$$

**Theorem 3.** *Assume (A1). Let  $\epsilon > 0$ ,  $\beta_0 > 0$  and  $\eta > 0$ . For any  $\alpha \in (0, 1]$ , there exists  $K_{\epsilon, \alpha, \beta_0, \eta} < \infty$  such that, for any  $\Sigma, \Sigma' \in \mathcal{M}_+^L(d, \epsilon)$  and any  $\beta, \beta' \in \mathcal{K}_{\beta_0, \eta}$ , it holds that*

$$D_{\mathbf{V}_{\beta_0}^\alpha}[(\Sigma, \beta), (\Sigma', \beta')] \leq K_{\epsilon, \alpha, \beta_0, \eta} \{|\beta - \beta'| + |\Sigma - \Sigma'|\}.$$

**3.3. Strong law of large numbers.** We can state an ergodicity result for the adaptive parallel tempering algorithm, given the step size sequences satisfy the following natural condition.

**(A2)** Assume that the step sizes  $\{\gamma_{n,i}, n \in \mathbb{N}\}$ ,  $i = 1, 2, 3$  defined in (14), (18), and (20) are non-negative and satisfy following conditions

- (i) For  $i = 1, 2, 3$ ,  $\sum_{n \geq 1} \gamma_{n,i} = \infty$  and  $\sum_{n \geq 1} n^{-1} \gamma_{n,i} < \infty$
- (ii)  $\sup_{n \in \mathbb{N}} \gamma_{n,2} \leq 1$

*Remark 4.* It is easy to see that  $\gamma_{n,i} = c_i(n+1)^{-\xi_i}$  with some  $c_1, c_3 \in (0, \infty)$  and  $c_2 \in (0, 1]$  and  $\xi_1, \xi_2, \xi_3 \in (0, 1]$  satisfy (A2).

**Theorem 5.** *Assume (A1) - (A2) and  $\mathbb{E}[\mathbf{V}_{\beta_0}(\mathbf{X}_0)] < \infty$ . Then, for any function  $f : \mathbf{X}^L \rightarrow \mathbb{R}$  such that  $\|f\|_{\mathbf{V}_{\beta_0}^\alpha} < \infty$  for some  $\alpha \in (0, 1)$  and given  $\lim_{n \rightarrow \infty} \pi_{\beta_n}(f)$  exists, we have*

$$\frac{1}{n} \sum_{i=1}^n f(\mathbf{X}_i) \longrightarrow \lim_{n \rightarrow \infty} \pi_{\beta_n}(f) \quad a.s.$$

*Remark 6.* In practice, one is usually only interested in integrating with respect to  $\pi$ , which means functions  $f$  depending only on the first coordinate, that is,  $f(x^{(1:L)}) = f(x^{(1)})$ . In this case, the limit condition is trivial, because  $\pi_{\beta_n}(f) = \pi(f)$  for all  $n \in \mathbb{N}$ .

**3.4. Convergence of temperature adaptation.** The strong law of large numbers (Theorem 5) does not require the convergence of the inverse temperatures, if only the coolest chain  $x^{(1)}$  is involved (Remark 6). It is, however, important to work out the convergence of the adaptation, because then we know what to expect on the asymptotic behaviour of the algorithm. Having the convergence, it is also possible to establish central limit theorems [1]; however, we do not pursue it here.

We denote the associated mean field of the stochastic approximation procedure (14) by

$$\mathbf{h}(\boldsymbol{\rho}) \stackrel{\text{def}}{=} [h^{(1)}(\rho^{(1)}), \dots, h^{(L-1)}(\rho^{(1)}), \dots, \rho^{(L-1)}],$$

where

$$h^{(\ell)}(\rho^{(1)}, \dots, \rho^{(\ell)}) \stackrel{\text{def}}{=} \int H^{(\ell)}(\rho^{(1)}, \dots, \rho^{(\ell)}, \mathbf{x}) \pi_{\beta}(\mathrm{d}\mathbf{x}).$$

We may write

$$h^{(\ell)}(\boldsymbol{\rho}) = \iint \varpi_{\beta}^{(\ell)}(x^{(\ell)}, x^{(\ell+1)}) \frac{\pi^{\beta^{(\ell)}}(\mathrm{d}x^{(\ell)})}{Z(\beta^{(\ell)})} \frac{\pi^{\beta^{(\ell+1)}}(\mathrm{d}x^{(\ell+1)})}{Z(\beta^{(\ell+1)})} - \alpha^*,$$

where  $Z(\beta)$  is the normalising constant defined in (2).

The following result establishes the existence and uniqueness of the stable point of the adaptation. In words, the following result implies that there exist unique temperatures so that the mean rate of accepting proposed swaps is  $\alpha^*$ .

**Proposition 7.** *Assume (A1). Then, there exists a unique  $\hat{\boldsymbol{\rho}} \stackrel{\text{def}}{=} [\hat{\rho}^{(1)}, \dots, \hat{\rho}^{(L-1)}]$  solution of the system of equations  $h^{(\ell)}(\rho^{(1)}, \dots, \rho^{(\ell)}) = 0$ ,  $\ell \in \{1, \dots, L-1\}$ .*

*Remark 8.* In Proposition Proposition 7, it is sufficient to assume that the support of  $\pi$  has infinite Lebesgue measure and that  $\int \pi^{\kappa}(x) \mathrm{d}x < \infty$  for all  $0 < \kappa \leq 1$ ; see Lemma 26.

*Remark 9.* In case the support of  $\pi$  has a finite Lebesgue measure, it is not difficult to show that for a sufficiently large number of levels  $L \geq L_0$  there is no solution  $\hat{\boldsymbol{\rho}}$ . On the contrary, in formal terms,  $\hat{\rho}^{(\ell)} = \infty$  for  $\ell \geq L_0$ , so that the corresponding inverse temperatures  $\hat{\beta}^{(\ell)} = 0$  for  $\ell \geq L_0 + 1$ . For our algorithm, this would imply that it simulates asymptotically  $\pi^0/Z(0)$ , the uniform distribution on the support of  $\pi$ , with the levels  $\ell \geq L_0 + 1$ .

For the convergence of the temperature adaptation, we require more stringent conditions on the step size sequence.

**(A3)** Assume that step sizes  $\{\gamma_{n,i}, n \in \mathbb{N}\}$  defined in (14),(18),(19) and (20) are non-negative and satisfy following conditions

- (i)  $\sum_{n \geq 1} \gamma_{n,i} = \infty$ ,  $\sum_{n \geq 1} \gamma_{n,1}^2 < \infty$ , and  $\sum_{n \geq 1} \gamma_{n,1} \gamma_{n,j} < \infty$ ,  $j = 2, 3$ .
- (ii)  $\sup_{n \in \mathbb{N}} \gamma_{n,2} \leq 1$
- (iii)  $\sum_{n \geq 1} |\gamma_{n+1,1} - \gamma_{n,1}| < \infty$

It is easy to check that the sequences introduced in Remark 4 satisfy **(A3)** if we assume in addition that  $\xi_1, \xi_2, \xi_3 \in (1/2, 1]$ .

**Theorem 10.** Assume **(A1)**–**(A3)**,  $\mathbb{E}V_{\beta_0}(\mathbf{X}_0) < \infty$ . In addition for all  $\ell = 1, \dots, L-1$  we assume that  $\underline{\rho} < \hat{\rho}^{(\ell)} < \bar{\rho}$ , where  $\hat{\rho}$  is given by Proposition 7. Then

$$\lim_{n \rightarrow \infty} \boldsymbol{\rho}_n = \hat{\boldsymbol{\rho}} \quad a.s..$$

#### 4. EXPERIMENTS

We consider two different type of examples: mixture of Gaussians in Section 4.1 and a challenging spatial imaging example in Section 4.2. In all the experiments, we use the step size sequences  $\gamma_{n,\cdot} = (n+1)^{-0.6}$ , except for RAM adaptation, where  $\gamma_{n,2} = \min\{0.9, d \cdot (n+1)^{-0.6}\}$  (see [33] for a discussion). We did not observe numerical instability issues, so the adaptations were not enforced to be stable by projections. We used the following initial values for the adapted parameters: temperature difference  $\rho_0^{(\ell)} = 1$ , covariances  $\Sigma_0^{(\ell)} = I$  and scalings  $\theta_0^{(\ell)} = 1$ .

**4.1. Mixture of Gaussians.** We consider first a well-known two-dimensional mixture of Gaussians example [e.g. 7, 19]. The example consists of 20 mixture components with means in  $[0, 10]^2$  and each component has a diagonal covariance  $\sigma^2 I$ , with  $\sigma^2 = 0.01$ . Figure 1 shows an example of the points simulated by our parallel tempering algorithm in this example, when we use  $L = 5$  energy levels and the default (covariance) adaptation to adjust the random walk proposals. Figure 2 shows the convergence of the temperature parameters in the same example.

We computed estimates of the means and the squares of the coordinates with  $N = 5000$  iterations of which 2500 burn-in, and show the mean and standard deviation (in parenthesis) over 100 runs of our parallel tempering algorithm in Table 1. We considered three different random-walk adaptations: the default adaptation in (18)–(20) (Cov), with common mean and covariance estimators as defined in (23)–(24) (Cov(g)) and the RAM update defined in (25). Table 1 shows the results in the same form as [7, Table 3] to allow easy comparison. When comparing with [7], our results show smaller deviation than the unadapted parallel tempering, but bigger deviation than their samplers including also equi-energy moves. We remark that we did not adjust

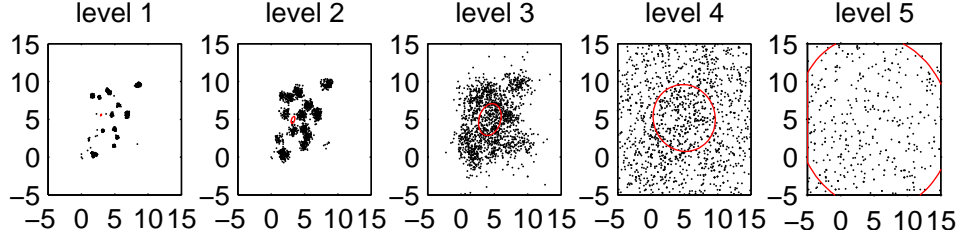


FIGURE 1. Simulated points of the tempered distributions over 5000 iterations. The random-walk proposal is illustrated as an ellipsoid.

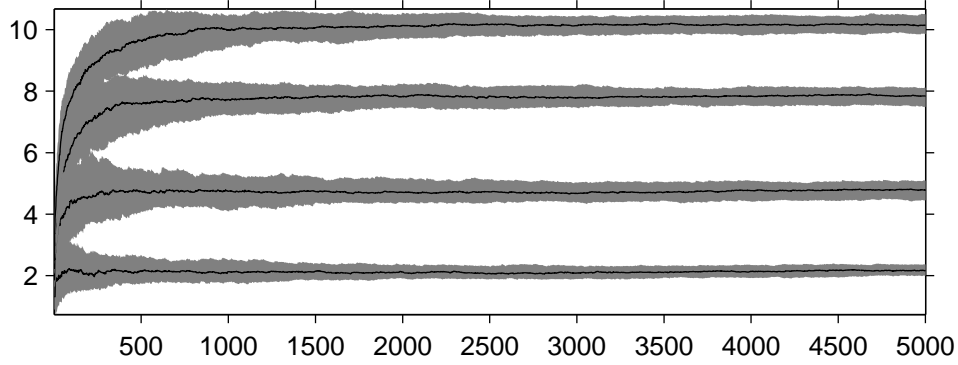


FIGURE 2. Convergence of the log-temperatures in the mixture example. The 10%–90% quantiles over 100 runs of the algorithm are drawn in grey and the black line shows the median.

TABLE 1. The test of [7], Table 3 case (A) with  $L = 5$  temperature levels, 5000 iterations and 2500 burn-in.

	$\mathbb{E}[X_1]$	$\mathbb{E}[X_2]$	$\mathbb{E}[X_1^2]$	$\mathbb{E}[X_2^2]$
True value	4.478	4.905	25.605	33.920
Cov	4.469 (0.588)	4.950 (0.813)	25.329 (5.639)	34.209 (8.106)
Cov(g)	4.389 (0.537)	4.803 (0.692)	24.677 (5.411)	32.865 (6.660)
RAM	4.530 (0.524)	4.946 (0.811)	26.111 (5.308)	34.331 (8.292)

our algorithm at all for the example, but let the adaptation take care of that. There are no significant differences between the random-walk adaptation algorithms.

When looking the simulated points in Figure 1, it is clear that three temperature levels is enough to allow good mixing in the above example. We repeated the example with  $L = 3$  energy levels, and increased the number of iterations to  $N = 8333$  in order to have a comparable

TABLE 2. The test of [7], Table 3 case (A) with  $L = 3$  temperature levels, 8333 iterations and 4167 burn-in.

	$\mathbb{E}[X_1]$	$\mathbb{E}[X_2]$	$\mathbb{E}[X_1^2]$	$\mathbb{E}[X_2^2]$
True value	4.478	4.905	25.605	33.920
Cov	4.480 (0.416)	4.957 (0.571)	25.542 (4.164)	34.420 (5.669)
Cov(g)	4.488 (0.422)	4.884 (0.551)	25.719 (4.190)	33.520 (5.476)
RAM	4.490 (0.407)	4.881 (0.541)	25.667 (4.281)	33.622 (5.631)

TABLE 3. The root mean square errors in the modified mixture example with  $\sigma^2 = 0.001$ ,  $d = 8$  and  $L = 8$ .

$N$	Est.	Cov	Cov(g)	RAM
$10k$	$\mathbb{E}[X]$	3.080	2.245	1.660
	$\mathbb{E}[ X ^2]$	27.577	20.426	16.428
$20k$	$\mathbb{E}[X]$	1.788	1.580	1.429
	$\mathbb{E}[ X ^2]$	18.577	15.712	14.475
$40k$	$\mathbb{E}[X]$	1.439	1.267	0.952
	$\mathbb{E}[ X ^2]$	15.471	12.769	9.364
$80k$	$\mathbb{E}[X]$	1.257	0.975	0.698
	$\mathbb{E}[ X ^2]$	13.017	9.414	6.981
$160k$	$\mathbb{E}[X]$	1.096	0.680	0.508
	$\mathbb{E}[ X ^2]$	11.093	7.038	5.122

computational cost. The summary of the results in Table 2 indicates increased accuracy than with  $L = 5$  levels, and the accuracy comes close to the results reported in [7] for samplers with equi-energy moves.

We considered also a more difficult modification of the mixture example above. We decreased the variance of the mixture components to  $\sigma^2 = 0.001$  and increased the dimension to  $d = 8$ . The mixture means of the added coordinates were all zero. We ran our adaptive parallel tempering algorithm in this case with  $L = 8$  temperature levels; Table 3 summarises the results with different number of iterations. In all the cases, the first half of the iterations were burn-in. In this scenario, the different random-walk adaptation algorithms have slightly different behaviour. Particularly, the common mean and covariance estimates (Cov(g)) seem to improve over the separate covariances (Cov). The RAM approach seems to provide the most accurate results. However, we believe that this is probably due to the special properties of the example, namely the fact that all the mixture components have a common covariance, and the RAM converges close to this in the first level; see the discussion in [33].

**4.2. Spatial imaging.** As another example, we consider identifying ice floes from polar satellite images as described by Banfield and Raftery [6]. The image under consideration is a 200 by 200 gray-scale satellite image, and we focus on the same 40 by 40 subregion as in [8]. The goal is to identify the presence and position of polar ice floes. Towards this goal, Higdon [14] employs a Bayesian model with an Ising model prior and following posterior distribution on  $\{0, 1\}^{40 \times 40}$ ,

$$\log(\pi(x|y)) \propto \sum_{1 \leq i, j \leq 40} \alpha \mathbb{1}\{x_{i,j} = y_{i,j}\} + \sum_{(i,j) \sim (i',j')} \beta \mathbb{1}\{x_{i,j} = x_{i',j'}\},$$

where neighbourhood relation ( $\sim$ ) is defined as vertical, horizontal and diagonal adjacencies of each pixel. Posterior distribution favours  $x$  which are similar to original image  $y$  (first term) and for which the neighbouring pixels are equal (second term).

In [14] and [8], the authors observed that standard MCMC algorithms which propose to flip one pixel at a time fail to explore the modes of the posterior. There are, however, some advantages of using such an algorithm, given we can overcome the difficulty in mixing between the modes. Specifically, in order to compute (the log-difference of) the unnormalised density values, we need only to explore the neighbourhoods of the pixels that have changed. Therefore, the proposal with one pixel flip at a time has a low computational cost. Moreover, such an algorithm is easy to implement.

We used our parallel tempering algorithm with the above mentioned proposal with  $L = 10$  temperature levels to simulate the posterior of this model with parameters  $\alpha = 1$  and  $\beta = 0.7$ . We ran 100 replications of  $N = 100000$  iterations of the algorithm. Obtained result are shown in Figure 3 is similar to [14] and [8]. We emphasize again that our algorithm provided good results without any prior tuning.

## 5. PROOFS

**5.1. Proof of Theorem 1.** The proof follows by arguments in the literature that guarantee a geometric ergodicity for the individual random-walk Metropolis kernels, and by observing that the swap kernel is invariant under permutation-invariant functions.

We start with an easy lemma showing that a drift in cooler chain implies a drift in the higher-temperature chain.

**Lemma 11.** *Consider the drift function  $W \stackrel{\text{def}}{=} c\pi^{-\kappa}$  for some positive constants  $\kappa$  and  $c$ . Then, for any  $\Sigma \in \mathcal{M}_+(d)$ ,*

$$\beta \leq \beta' \implies M_{(\Sigma, \beta')} W(x) \leq M_{(\Sigma, \beta)} W(x), \quad \text{for all } x \in \mathsf{X}.$$

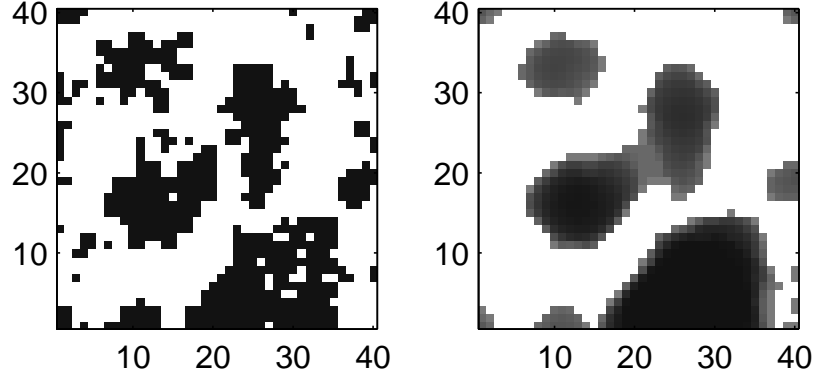


FIGURE 3. Spatial model example: original image (left), posterior estimate based on 100 replications of adaptive parallel tempering (right)

*Proof.* We write

$$\begin{aligned} \frac{M_{(\Sigma, \beta)} W(x)}{W(x)} &= \int_{\{y : \pi(y) \geq \pi(x)\}} \left( \frac{\pi(x)}{\pi(y)} \right)^\kappa q_\Sigma(x - y) dy \\ &\quad + \int_{\{y : \pi(y) < \pi(x)\}} \left[ 1 - \left( \frac{\pi(y)}{\pi(x)} \right)^\beta + \left( \frac{\pi(y)}{\pi(x)} \right)^{\beta - \kappa} \right] q_\Sigma(x - y) dy \end{aligned}$$

First term is independent on  $\beta$ , since  $\beta \mapsto 1 - a^\beta + a^{\beta - \kappa}$  for  $a \in [0, 1]$  is non-increasing the second term is also non-increasing with respect to  $\beta$ .  $\square$

To control the ergodicity of each individual random-walk Metropolis sampler, it is required to have a control on the minorisation and drift constants for the kernels  $M_{(\Sigma, \beta)}$ . The following proposition provides such explicit control.

**Lemma 12.** *Assume (A1). Let  $\varepsilon > 0$  and  $\beta > 0$ . There exist  $\lambda_{\varepsilon, \beta} \in [0, 1)$  and  $b_{\varepsilon, \beta} < \infty$  such that for any  $\Sigma \in \mathcal{M}_+(d, \varepsilon)$ , we get*

$$(38) \quad M_{(\Sigma, \beta)} V_\beta \leq \lambda_{\varepsilon, \beta} V_\beta + b_{\varepsilon, \beta},$$

where  $V_\beta \stackrel{\text{def}}{=} (\pi / \|\pi\|_\infty)^{-\beta/2}$ , where  $\|\pi\|_\infty = \sup_{x \in \mathbf{X}} \pi(x)$ .

*Proof.* It is easily seen that if the target distribution is super-exponential in the tails (A1), then all the tempered versions  $\pi^\beta / Z(\beta)$ , where the normalising constant  $Z(\beta)$  is defined in (2), satisfy (A1) as well.

The result then follows from Andrieu and Moulines [1, Proposition 12].  $\square$

**Proposition 13.** *Assume (A1) and let  $\epsilon > 0$  and  $\beta_0 > 0$ . Then, there exists  $\lambda_{\epsilon, \beta_0} < 1$ , and  $b_{\epsilon, \beta_0} < \infty$  such that, for all  $\Sigma \in \mathcal{M}_+^L(d, \epsilon)$  and  $\beta \in \mathcal{K}_{\beta_0}$ ,*

$$(39) \quad \mathbf{M}_{(\Sigma, \beta)} \mathbf{V}_{\beta_0} \leq \lambda_{\epsilon, \beta_0} \mathbf{V}_{\beta_0} + b_{\epsilon, \beta_0},$$

$$(40) \quad \mathbf{S}_{\beta} \mathbf{V}_{\beta_0} = \mathbf{V}_{\beta_0},$$

$$(41) \quad \mathbf{P}_{(\Sigma, \beta)} \mathbf{V}_{\beta_0} \leq \lambda_{\epsilon, \beta_0} \mathbf{V}_{\beta_0} + b_{\epsilon, \beta_0},$$

*Proof.* By Lemma 11, since  $\beta \in \mathcal{K}_{\beta_0}$ , we get

$$(42) \quad \begin{aligned} \mathbf{M}_{(\Sigma, \beta)} \mathbf{V}_{\beta_0}(x^{(1:L)}) &= \sum_{\ell=1}^L M_{(\Sigma^{(\ell)}, \beta^{(\ell)})} V_{\beta_0}(x^{(\ell)}) \\ &\leq \sum_{\ell=1}^L M_{(\Sigma^{(\ell)}, \beta_0)} V_{\beta_0}(x^{(\ell)}). \end{aligned}$$

Then, by Lemma 12, since  $\Sigma \in \mathcal{M}_+(d, \epsilon)$ , it holds

$$\sum_{\ell=1}^L M_{(\Sigma^{(\ell)}, \beta_0)} V_{\beta_0}(x^{(\ell)}) \leq \lambda_{\epsilon, \beta_0} \sum_{\ell=1}^L V_{\beta_0}(x^{(\ell)}) + Lb_{\epsilon, \beta_0}.$$

Thanks to the definition of the swapping kernel (8)–(10), for any positive measurable function  $F : \mathbf{X}^L \rightarrow \mathbb{R}_+$  which is invariant by permutation<sup>1</sup>, we get  $\mathbf{S}_{\beta} F(x^{(1:L)}) = F(x^{(1:L)})$ . Since the drift function  $\mathbf{V}_{\beta_0}$  defined in (30) is invariant by permutation we obtain

$$\begin{aligned} \mathbf{S}_{\beta} [\lambda_{\epsilon, \beta_0} \mathbf{V}_{\beta_0}(x^{(1:L)}) + Lb_{\epsilon, \beta_0}] &= \lambda_{\epsilon, \beta_0} \mathbf{S}_{\beta} \mathbf{V}_{\beta_0}(x^{(1:L)}) + Lb_{\epsilon, \beta_0} \\ &= \lambda_{\epsilon, \beta_0} \mathbf{V}_{\beta_0}(x^{(1:L)}) + Lb_{\epsilon, \beta_0}. \quad \square \end{aligned}$$

**Proposition 14.** *Assume (A1). Let  $\epsilon > 0$ ,  $\beta_0 > 0$  and  $r > 1$ , and consider the level set  $\mathbf{C}_r \stackrel{\text{def}}{=} \{\mathbf{x} \in \mathbf{X}^L : \mathbf{V}_{\beta_0}(\mathbf{x}) \leq r\}$ . There exists a constant  $\delta_{r, \epsilon, \beta_0} > 0$  such that for all  $\Sigma \in \mathcal{M}_+(d, \epsilon)$  and  $\beta \in \mathcal{K}_{\beta_0}$ , the set  $\mathbf{C}_r$  is a  $(1, \delta_{r, \epsilon, \beta_0}, \nu_r)$ -small set for  $\mathbf{P}_{(\Sigma, \beta)}$ , that is,*

$$(43) \quad \mathbf{P}_{(\Sigma, \beta)}(\mathbf{x}, \cdot) \geq \delta_{r, \epsilon, \beta_0} \nu_{r, \epsilon, \beta_0}(\cdot) \quad \mathbf{x} \in \mathbf{C}_r,$$

where  $\nu_r(\cdot) = \lambda^{\text{Leb}}(\cdot \cap \mathbf{C}_r) / \lambda^{\text{Leb}}(\mathbf{C}_r)$  is a probability measure on  $\mathbf{C}_r$  and  $\lambda^{\text{Leb}}$  stands for the Lebesgue measure.

*Proof.* It is easy to see that the set  $\mathbf{C}_r$  is absorbing for  $\mathbf{S}_{\beta}$  because  $\mathbf{S}_{\beta} V_{\beta}(x) = V_{\beta}(x)$  as observed in the proof of Proposition 14, implying  $\mathbf{S}_{\beta}(x, \mathbf{C}_r) = 1$  for all  $x \in \mathbf{C}_r$ . Hence for  $x \in \mathbf{C}_r$

$$(44) \quad \mathbf{P}_{(\Sigma, \beta)}(\mathbf{x}, \mathbf{A}) \geq \int_{\mathbf{C}_r \cap \mathbf{A}} \prod_{\ell=1}^L \left( 1 \wedge \frac{\pi(y^{(\ell)})}{\pi(x^{(\ell)})} \right)^{\beta^{(\ell)}} q_{\Sigma^{(\ell)}}(y^{(\ell)} - x^{(\ell)}) dy^{(1:L)},$$

---

<sup>1</sup> $F(x^{(1:L)}) = F(x^{(\sigma(1):\sigma(L))})$  for any  $(x^{(1:L)}) \in \mathbf{X}^L$  and any permutation  $\sigma$  over the set  $\{1, \dots, L\}$



where  $q_\Sigma$  is the multivariate Gaussian density with zero mean and covariance  $\Sigma$ . Since the set  $\mathbf{C}_r$  is compact and  $\Sigma \in \mathcal{M}_+^L(\beta_0, \epsilon)$ , there exists a constant  $C_{r,\epsilon,\beta_0} > 0$  such that for any  $\ell = 1, \dots, L$

$$\inf_{(x,y) \in \mathbf{C}_r \times \mathbf{C}_r} q^{(\ell)}(y^{(\ell)} - x^{(\ell)}) \geq C_{r,\epsilon,\beta_0}.$$

Therefore, by (44) and since  $\beta^{(\ell)} \in (0, 1]$ , we obtain for  $x \in \mathbf{C}_r$

$$\mathbf{P}_{(\Sigma,\beta)}(\mathbf{x}, \mathbf{A}) \geq C_{r,\epsilon,\beta_0}^L \int_{\mathbf{C}_r \cap \mathbf{A}} \prod_{\ell=1}^L \left( 1 \wedge \frac{\pi(y^{(\ell)})}{\pi(x^{(\ell)})} \right) dy^{(1:L)}.$$

If  $y = (y^{(1:L)}) \in \mathbf{C}_r$ , we get  $(\pi(y^{(\ell)})/\|\pi\|_\infty)^{-\beta_0} \leq r/L$  for all  $\ell \in \{1, \dots, L\}$ , which implies that  $(L/r)^{2/\beta_0} \|\pi\|_\infty \leq \pi(y)$ . Hence, for all  $(x, y) \in \mathbf{C}_r \times \mathbf{C}_r$ ,  $\pi(y)/\pi(x) \leq (L/r)^{2/\beta_0}$ , showing that

$$\mathbf{P}_{(\Sigma,\beta)}(\mathbf{x}, \mathbf{A}) \geq C_{r,\epsilon,\beta_0}^L [(1 \wedge (L/r))^{2/\beta_0}]^L \lambda^{\text{Leb}}(\mathbf{C}_r) \lambda_{\mathbf{C}_r}^{\text{Leb}}(\mathbf{C}_r \cap \mathbf{A}),$$

where  $\lambda_{\mathbf{C}_r}^{\text{Leb}}(\cdot) = \lambda^{\text{Leb}}(\mathbf{C}_r \cap \cdot)/\lambda^{\text{Leb}}(\mathbf{C}_r)$ .  $\square$

*Proof of Theorem 1.* Choose a sufficiently large  $r > 1$  so that there exists a  $\tilde{\lambda}_{\epsilon,\beta_0} < 1$  such that

$$\mathbf{P}_{(\Sigma,\beta)} \mathbf{V}_{\beta_0}(x) \leq \lambda_{\epsilon,\beta_0} \mathbf{V}_{\beta_0}(x) + \mathbb{I}\{x \notin C\} b_{\epsilon,\beta_0}$$

by Proposition 13, where  $\mathbf{C}_r$  is defined in Proposition 14. This drift inequality, with the minorisation inequality in Proposition 14 imply  $V$ -uniform ergodicity (33) with constants depending only on  $\tilde{\lambda}_{\epsilon,\beta_0} < 1$ ,  $b_{\epsilon,\beta_0}$  and  $\delta_{r,\epsilon,\beta_0}$  [e.g. 23].  $\square$

**5.2. Proof of Theorem 3.** We preface the proof of this Theorem by several technical lemmas.

**Lemma 15.** *For all  $(\beta, \beta') \in (0, 1)^2$  we have*

$$\sup_{z \in [0,1]} |z^\beta - z^{\beta'}| \leq \frac{1}{\max\{\beta, \beta'\}} |\beta' - \beta|.$$

*Proof.* Without loss of generality, assume that  $0 < \beta < \beta' < 1$ . The function  $w : [0, 1] \rightarrow \mathbb{R}$  defined as  $w(z) = z^\beta - z^{\beta'}$  is continuous, non-negative and  $w(0) = w(1) = 0$ . Therefore, the maximum of this function is obtained inside the interval  $(0, 1)$ . By computing the derivative  $w'(z) = \beta z^{\beta-1} - \beta' z^{\beta'-1}$  and setting  $w'(z) = 0$ , we find the maximum at  $z = (\frac{\beta}{\beta'})^{1/(\beta'-\beta)}$ , so

$$\sup_{z \in [0,1]} |z^\beta - z^{\beta'}| = \left(\frac{\beta}{\beta'}\right)^{\frac{\beta}{\beta'-\beta}} \left(1 - \frac{\beta}{\beta'}\right) \leq \left(1 - \frac{\beta}{\beta'}\right) = \frac{1}{\beta'}(\beta' - \beta). \quad \square$$

**Lemma 16.** *Set  $\beta_0 \in (0, 1]$ . There exists a constant  $K_{\beta_0} < \infty$  such that for any  $(\beta, \beta') \in [\beta_0, 1]^2$ , any covariance matrix  $\Sigma \in \mathcal{M}_+(d)$ ,*

$$(45) \quad \int_{\mathbf{X}} (V_{\beta_0}(y) + V_{\beta_0}(x)) |\alpha_{\beta}(x, y) - \alpha_{\beta'}(x, y)| q_{\Sigma}(y - x) \lambda^{\text{Leb}}(dy) \leq K_{\beta_0} |\beta - \beta'| V_{\beta_0}(x).$$

*In addition, for any measurable function  $g$  with  $\|g\|_{V_{\beta_0}} \leq 1$ ,*

$$(46) \quad |M_{(\Sigma, \beta)} g(x) - M_{(\Sigma, \beta')} g(x)| \leq K_{\beta_0} |\beta - \beta'| V_{\beta_0}(x).$$

*Proof.* Without loss of generality, we assume that  $\beta < \beta'$ . Recall that  $V_{\beta_0}(y) \propto \pi^{-\beta_0/2}(y)$ . Note that

$$\begin{aligned} & \int_{\mathbf{X}} V_{\beta_0}(y) |\alpha_{\beta}(x, y) - \alpha_{\beta'}(x, y)| q_{\Sigma}(y - x) dy \\ &= V_{\beta_0}(x) \int_{\mathbf{R}_x} \left| \left( \frac{\pi(y)}{\pi(x)} \right)^{\beta - \beta_0/2} - \left( \frac{\pi(y)}{\pi(x)} \right)^{\beta' - \beta_0/2} \right| q_{\Sigma}(y - x) dy, \end{aligned}$$

where  $\mathbf{R}_x := \{y \in \mathbf{X} : \pi(y) < \pi(x)\}$ , and

$$\begin{aligned} & V_{\beta_0}(x) \int_{\mathbf{X}} |\alpha_{\beta}(x, y) - \alpha_{\beta'}(x, y)| q_{\Sigma}(y - x) dy \\ &= V_{\beta_0}(x) \int_{\mathbf{R}_x} \left| \left( \frac{\pi(y)}{\pi(x)} \right)^{\beta} - \left( \frac{\pi(y)}{\pi(x)} \right)^{\beta'} \right| q_{\Sigma}(y - x) dy. \end{aligned}$$

Using Lemma 15, we get

$$\int_{\mathbf{R}_x} \left| \left( \frac{\pi(y)}{\pi(x)} \right)^{\beta - \beta_0/2} - \left( \frac{\pi(y)}{\pi(x)} \right)^{\beta' - \beta_0/2} \right| q_{\Sigma}(y - x) dy \leq \frac{1}{\beta - \beta_0/2} |\beta' - \beta|,$$

and similarly

$$\int_{\mathbf{R}_x} \left| \left( \frac{\pi(y)}{\pi(x)} \right)^{\beta} - \left( \frac{\pi(y)}{\pi(x)} \right)^{\beta'} \right| q_{\Sigma}(y - x) dy \leq \frac{1}{\beta} |\beta' - \beta|,$$

which concludes the proof of (45).

We consider now (46). Note that

$$\begin{aligned} (47) \quad & |M_{(\Sigma, \beta)} g(x) - M_{(\Sigma, \beta')} g(x)| \\ &= \left| \int_{\mathbf{X}} (g(y) - g(x)) (\alpha_{\beta}(x, y) - \alpha_{\beta'}(x, y)) q_{\Sigma}(y - x) dy \right| \\ &\leq \int_{\mathbf{X}} |g(y) - g(x)| |\alpha_{\beta}(x, y) - \alpha_{\beta'}(x, y)| q_{\Sigma}(y - x) dy \\ &\leq \int_{\mathbf{X}} (V_{\beta_0}(y) + V_{\beta_0}(x)) |\alpha_{\beta}(x, y) - \alpha_{\beta'}(x, y)| q_{\Sigma}(y - x) dy \end{aligned}$$

and we conclude using (45).  $\square$

The following result is a restatement of [1, Proposition 12 and (the proof of) Lemma 13].

**Lemma 17.** *For any  $\epsilon > 0$  there exists  $K_\epsilon < \infty$  such that for any  $(\Sigma, \Sigma') \in \mathcal{M}_+(d, \epsilon)$ ,  $\beta \in [0, 1]$ , and function  $g$  with  $\|g\|_{V_{\beta_0}} \leq 1$ , we have*

$$|M_{(\Sigma, \beta)}g(x) - M_{(\Sigma', \beta)}g(x)| \leq K_\epsilon |\Sigma - \Sigma'| V_{\beta_0}(x).$$

In addition there exists  $K_q < \infty$  such that

$$(48) \quad \int_{\mathbf{x}} |q_\Sigma(z) - q_{\Sigma'}(z)| dz \leq K_q |\Sigma - \Sigma'|.$$

Lemma 17 can be generalised also for non-Gaussian proposal distributions, including the multivariate Student [31, Appendix B].

Now we show the local Lipschitz-continuity of the mapping  $(\Sigma, \beta) \mapsto \mathbf{M}_{(\Sigma, \beta)}$ .

**Proposition 18.** *Assume (A1). Let  $\epsilon > 0$  and  $\beta_0 > 0$ , and  $\eta > 0$ . There exists a constant  $K_{\epsilon, \beta_0, \eta} < \infty$  such that for all  $\Sigma, \Sigma' \in \mathcal{M}_+^L(d, \epsilon)$ ,  $\beta, \beta' \in \mathcal{K}_{\beta_0, \eta}$ , and functions  $g$  with  $\|g\|_{\mathbf{V}_{\beta_0}} \leq 1$ , we have*

$$|\mathbf{M}_{(\Sigma, \beta)}g(\mathbf{x}) - \mathbf{M}_{(\Sigma', \beta')}g(\mathbf{x})| \leq K_{\epsilon, \beta_0, \eta} \{|\beta - \beta'| + |\Sigma - \Sigma'|\} \mathbf{V}_{\beta_0}(\mathbf{x}).$$

*Remark 19.* In this proof, it is possible to set  $\eta = 0$ . The use of  $\eta > 0$  is required in the proof of continuity.

*Proof.* We may write

$$(49) \quad \begin{aligned} |\mathbf{M}_{(\Sigma, \beta)}g(\mathbf{x}) - \mathbf{M}_{(\Sigma', \beta')}g(\mathbf{x})| &\leq |\mathbf{M}_{(\Sigma, \beta)}g(\mathbf{x}) - \mathbf{M}_{(\Sigma, \beta')}g(\mathbf{x})| \\ &\quad + |\mathbf{M}_{(\Sigma, \beta')}g(\mathbf{x}) - \mathbf{M}_{(\Sigma', \beta')}g(\mathbf{x})| \stackrel{\text{def}}{=} R_1(\mathbf{x}) + R_2(\mathbf{x}). \end{aligned}$$

First, we consider  $R_1$ . We prove by induction that there exists a constant  $K_{k, \epsilon, \beta_0, \eta} < \infty$  such that, for all measurable  $g$  such that  $\|g\|_{\mathbf{V}_{\beta_0}} \leq 1$ ,

$$(50) \quad \left| \mathbf{M}_{(\Sigma, \beta)}^{(k)}g(\mathbf{x}) - \mathbf{M}_{(\Sigma, \beta')}^{(k)}g(\mathbf{x}) \right| \leq K_{k, \epsilon, \beta_0, \eta} |\beta - \beta'| \mathbf{V}_{\beta_0}(\mathbf{x}),$$

where  $\mathbf{M}_{(\Sigma, \beta)}^{(k)} \stackrel{\text{def}}{=} \prod_{1 \leq \ell \leq k} \check{M}_{(\Sigma^{(\ell)}, \beta^{(\ell)})}$ , and

$$\check{M}_{(\Sigma^{(\ell)}, \beta^{(\ell)})}(x^{(1:L)}, A_1 \times \cdots \times A_L) \stackrel{\text{def}}{=} M_{(\Sigma^{(\ell)}, \beta^{(\ell)})}(x^{(\ell)}, A_\ell) \prod_{i \neq \ell} \delta_{x^{(i)}}(A_i).$$

We first establish the result for  $k = 1$ . For any  $\ell \in \{1, \dots, L\}$ , we get

$$\begin{aligned}
& \left| \check{M}_{(\Sigma^{(\ell)}, \beta^{(\ell)})} g(\mathbf{x}) - \check{M}_{(\Sigma^{(\ell)}, \beta'^{(\ell)})} g(\mathbf{x}) \right| \\
& \leq \int_{\mathbf{x}} \left| g(x^{(1:\ell-1)}, y, x^{(\ell+1:L)}) - g(x^{(1:L)}) \right| \\
& \quad \left| \alpha_{\beta^{(\ell)}}(x^{(\ell)}, y) - \alpha_{\beta'^{(\ell)}}(x^{(\ell)}, y) \right| q_{\Sigma^{(\ell)}}(y - x^{(\ell)}) dy \\
& \leq \int_{\mathbf{x}} \left( V_{\beta_0}(y) + V_{\beta_0}(x^{(\ell)}) + 2 \sum_{k \neq \ell} V_{\beta_0}(x^{(k)}) \right) \\
& \quad \left| \alpha_{\beta^{(\ell)}}(x^{(\ell)}, y) - \alpha_{\beta'^{(\ell)}}(x^{(\ell)}, y) \right| q_{\Sigma^{(\ell)}}(y - x^{(\ell)}) dy
\end{aligned}$$

Applying (45), there exists  $K < \infty$  such that for any  $\ell \in \{1, \dots, L\}$ , we get

$$(51) \quad \left| \check{M}_{(\Sigma^{(\ell)}, \beta^{(\ell)})} g(\mathbf{x}) - \check{M}_{(\Sigma^{(\ell)}, \beta'^{(\ell)})} g(\mathbf{x}) \right| \leq K |\beta^{(\ell)} - \beta'^{(\ell)}| \sum_{k=1}^L V_{\beta_0}(x^{(k)})$$

Taking  $\ell = 1$  establishes (50) with  $k = 1$ . Assume now that (50) is satisfied for some  $k \in \{2, \dots, L-1\}$ . We have

$$\begin{aligned}
& \left| \mathbf{M}_{(\Sigma, \beta)}^{(k+1)} g(\mathbf{x}) - \mathbf{M}_{(\Sigma, \beta')}^{(k+1)} g(\mathbf{x}) \right| \\
& \leq \left| \left( \check{M}_{(\Sigma^{(k+1)}, \beta^{(k+1)})} - \check{M}_{(\Sigma^{(k+1)}, \beta'^{(k+1)})} \right) \mathbf{M}_{(\Sigma, \beta)}^{(k)} g(\mathbf{x}) \right| \\
& \quad + \left| \check{M}_{(\Sigma^{(k+1)}, \beta'^{(k+1)})} \left( \mathbf{M}_{(\Sigma, \beta)}^{(k)} - \mathbf{M}_{(\Sigma, \beta')}^{(k)} \right) g(\mathbf{x}) \right|
\end{aligned}$$

For any  $\|g\|_{\mathbf{V}_{\beta_0}} \leq 1$ , we have

$$\left| \mathbf{M}_{(\Sigma, \beta)}^{(k)} g(\mathbf{x}) \right| \leq \mathbf{M}_{(\Sigma, \beta)}^{(k)} \mathbf{V}_{\beta_0}(\mathbf{x}) = \sum_{\ell=1}^k M_{(\Sigma^{(\ell)}, \beta^{(\ell)})} V_{\beta_0}(x^{(\ell)}) + \sum_{\ell=k+1}^L V_{\beta_0}(x^{(\ell)}).$$

Lemma 12 implies that, for any  $\|g\|_{\mathbf{V}_{\beta_0}} \leq 1$ ,

$$\left| \mathbf{M}_{(\Sigma, \beta)}^{(k)} g(\mathbf{x}) \right| \leq \mathbf{V}_{\beta_0}(\mathbf{x}) + kb_{\epsilon, \beta_0},$$

showing that  $\|\mathbf{M}_{(\Sigma, \beta)}^{(k)} g\|_{\mathbf{V}_{\beta_0}} < \infty$ . Hence by (51) there exists a constant  $K_{k+1, \epsilon, \beta_0, \eta}^{(1)} < \infty$  such that

$$\begin{aligned}
& \left| \left( \check{M}_{(\Sigma^{(k+1)}, \beta^{(k+1)})} - \check{M}_{(\Sigma^{(k+1)}, \beta'^{(k+1)})} \right) \mathbf{M}_{(\Sigma, \beta)}^{(k)} g(\mathbf{x}) \right| \\
& \leq K_{k+1, \epsilon, \beta_0, \eta}^{(1)} |\beta^{(k+1)} - \beta'^{(k+1)}| \mathbf{V}_{\beta_0}(\mathbf{x}).
\end{aligned}$$

By the induction assumption (50) and Lemma 12 there exists  $K_{k+1,\epsilon,\beta_0,\eta}^{(2)} < \infty$  such that second term is bounded by

$$\begin{aligned} & \left| M_{(\Sigma^{(k+1)}, \beta'^{(k+1)})} (\mathbf{M}_{(\Sigma, \beta)}^{(k)} - \mathbf{M}_{(\Sigma, \beta')}^{(k)}) g(\mathbf{x}) \right| \\ & \leq K_{k,\epsilon,\beta_0,\eta} |\beta - \beta'| M_{(\Sigma^{(k+1)}, \beta'^{(k+1)})} \mathbf{V}_{\beta_0}(\mathbf{x}) \leq K_{k+1,\epsilon,\beta_0,\eta}^{(2)} |\beta - \beta'| \mathbf{V}_{\beta_0}(\mathbf{x}). \end{aligned}$$

This shows that (50) is satisfied for  $k+1$ . Carrying out the induction until  $k = L$ , there exists a constant  $K_{L,\epsilon,\beta_0,\eta} < \infty$  such that, for all  $\|g\|_{\mathbf{V}_{\beta_0}} \leq 1$ ,

$$(52) \quad R_1(x) = |\mathbf{M}_{(\Sigma, \beta)} g(\mathbf{x}) - \mathbf{M}_{(\Sigma, \beta')} g(\mathbf{x})| \leq K_{L,\epsilon,\beta_0,\eta} |\beta - \beta'| \mathbf{V}_{\beta_0}(\mathbf{x}),$$

where  $R_1$  is defined in (49).

Consider now  $R_2$ . It is easy to see that by (48) we obtain analogous formula to (51); that is, with the same temperatures  $\beta$  but different covariance matrices  $\Sigma$  and  $\Sigma'$ . The proof is concluded by using the same induction proof as for the term  $R_1$ .  $\square$

**Lemma 20.** *Let  $\beta_0 > 0$  and  $\eta > 0$ . Then, there exists a constant  $K_{\beta_0,\eta}$  such that, for any  $\beta, \beta' \in \mathcal{K}_{\beta_0,\eta}$  and for any measurable function  $g$  with  $\|g\|_{\mathbf{V}_{\beta_0}} \leq 1$ , it holds that*

$$|\mathbf{S}_\beta g(\mathbf{x}) - \mathbf{S}_{\beta'} g(\mathbf{x})| \leq K_{\beta_0,\eta} |\beta - \beta'| \mathbf{V}_{\beta_0}(\mathbf{x}).$$

*Proof.* Using the definition (8) of  $\mathbf{S}_\beta$ , we get

$$\begin{aligned} (53) \quad & |\mathbf{S}_\beta g(\mathbf{x}) - \mathbf{S}_{\beta'} g(\mathbf{x})| = \frac{1}{L-1} \left| \sum_{\ell=1}^{L-1} \left( \varpi_\beta^{(\ell)}(x^{(\ell)}, x^{(\ell+1)}) - \varpi_{\beta'}^{(\ell)}(x^{(\ell)}, x^{(\ell+1)}) \right) \right. \\ & \left. \times \left( g(x^{(1:\ell-1)}, x^{(\ell+1)}, x^{(\ell)}, \dots, x^{(L)}) - g(x^{(1:\ell-1)}, x^{(\ell)}, x^{(\ell+1)}, \dots, x^{(L)}) \right) \right|. \end{aligned}$$

By (9), it holds that  $\varpi_\beta^{(\ell)}(x^{(\ell)}, x^{(\ell+1)}) = \varpi_{\beta'}^{(\ell)}(x^{(\ell)}, x^{(\ell+1)}) = 1$  whenever  $\pi(x^{(\ell+1)}) \geq \pi(x^{(\ell)})$ . Therefore, using Lemma 15, we get that

$$\begin{aligned} (54) \quad & \left| \varpi_\beta^{(\ell)}(x^{(\ell)}, x^{(\ell+1)}) - \varpi_{\beta'}^{(\ell)}(x^{(\ell)}, x^{(\ell+1)}) \right| \\ & = \mathbb{I}_{\{\pi(x^{(\ell+1)}) \leq \pi(x^{(\ell)})\}} \left| \left( \frac{\pi(x^{(\ell+1)})}{\pi(x^{(\ell)})} \right)^{\beta^{(\ell)} - \beta^{(\ell+1)}} - \left( \frac{\pi(x^{(\ell+1)})}{\pi(x^{(\ell)})} \right)^{\beta'^{(\ell)} - \beta'^{(\ell+1)}} \right| \\ & \leq \frac{|\beta^{(\ell+1)} - \beta'^{(\ell+1)}| + |\beta^{(\ell)} - \beta'^{(\ell)}|}{(\beta^{(\ell)} - \beta^{(\ell+1)}) \wedge (\beta'^{(\ell)} - \beta'^{(\ell+1)})}. \end{aligned}$$

Since  $\beta, \beta' \in \mathcal{K}_{\beta_0,\eta}$ ,  $\max\{(\beta^{(\ell)} - \beta^{(\ell+1)}), (\beta'^{(\ell)} - \beta'^{(\ell+1)})\} \geq \eta$  for all  $\ell \in \{1, \dots, L\}$ . Because  $\|g\|_{\mathbf{V}_{\beta_0}} \leq 1$  and  $\mathbf{V}_{\beta_0}$  are invariant under

permutations, we have by (53) and (54)

$$|\mathbf{S}_\beta g(\mathbf{x}) - \mathbf{S}_{\beta'} g(\mathbf{x})| \leq \frac{4}{(L-1)\eta} \mathbf{V}_{\beta_0}(\mathbf{x}) \sum_{\ell=1}^L |\beta^{(\ell)} - \beta'^{(\ell)}|. \quad \square$$

Now we are ready to conclude with the continuity of the parallel tempering kernels.

*Proof of Theorem 3.* The definition (30) of  $\mathbf{V}_{\beta_0}$  implies that, for any  $\alpha \in (0, 1]$ ,

$$L^{\alpha-1} \sum_{i=1}^L V_{\alpha\beta_0}(x^{(i)}) \leq \left( \sum_{i=1}^L V_{\beta_0}(x^{(i)}) \right)^\alpha \leq \sum_{i=1}^L V_{\alpha\beta_0}(x^{(i)}),$$

showing that  $L^{\alpha-1} \mathbf{V}_{\alpha\beta_0}(\mathbf{x}) \leq \mathbf{V}_{\beta_0}^\alpha(\mathbf{x}) \leq \mathbf{V}_{\alpha\beta_0}(\mathbf{x})$ . Therefore, the norms  $\|\cdot\|_{\mathbf{V}_{\beta_0}^\alpha}$  and  $\|\cdot\|_{\mathbf{V}_{\alpha\beta_0}}$  are equivalent. It suffices to prove the results with  $\alpha = 1$ . Write

$$|\mathbf{P}_{(\Sigma, \beta)} g(\mathbf{x}) - \mathbf{P}_{(\Sigma', \beta')} g(\mathbf{x})| \leq T_1(\mathbf{x}) + T_2(\mathbf{x}),$$

where

$$\begin{aligned} T_1(\mathbf{x}) &\stackrel{\text{def}}{=} |\mathbf{S}_\beta (\mathbf{M}_{(\Sigma, \beta)} - \mathbf{M}_{(\Sigma', \beta')}) g(\mathbf{x})| \\ T_2(\mathbf{x}) &\stackrel{\text{def}}{=} |(\mathbf{S}_\beta - \mathbf{S}_{\beta'}) \mathbf{M}_{(\Sigma', \beta')} g(\mathbf{x})|. \end{aligned}$$

By Proposition 18, we obtain

$$\begin{aligned} T_1(\mathbf{x}) &\leq K_{\epsilon, \beta_0, \eta} \{|\beta - \beta'| + |\Sigma - \Sigma'|\} \mathbf{S}_\beta \mathbf{V}_{\beta_0}(\mathbf{x}) \\ &\leq K_{\epsilon, \beta_0, \eta} \{|\beta - \beta'| + |\Sigma - \Sigma'|\} \mathbf{V}_{\beta_0}(\mathbf{x}). \end{aligned}$$

By (39) of Lemma 12, we obtain that  $\|\mathbf{M}_{(\Sigma', \beta')} g\|_{\mathbf{V}_{\beta_0}} \leq C$ . Hence, by Lemma 20 we get that

$$T_2(\mathbf{x}) \leq CK_{\beta_0, \eta} |\beta - \beta'| \mathbf{V}_{\beta_0}(\mathbf{x}),$$

which concludes the proof.  $\square$

**5.3. Proof of Theorem 5.** We now turn into the proof of the strong law of large numbers. We start by gathering some known results and by technical lemmas.

**Lemma 21.** *Assume (A1) and that, in addition,  $\mathbb{E}[\mathbf{V}_{\beta_0}(\mathbf{X}_0)] < \infty$ . Then,*

$$\sup_{n \geq 1} \mathbb{E}[\mathbf{V}_{\beta_0}(\mathbf{X}_n)] < \infty,$$

where  $\mathbf{X}_n$  is the state of the adaptive parallel tempering algorithm defined in (4).

*Proof.* Under **(A1)**, by Proposition 13, for all  $n \in \mathbb{N}$  we have that

$$(55) \quad \mathbf{P}_{(\boldsymbol{\Sigma}_n, \boldsymbol{\beta}_n)} \mathbf{V}_{\beta_0} \leq \lambda_{\epsilon, \beta_0} \mathbf{V}_{\beta_0} + b.$$

Iterating (55), by (27) we obtain

$$\begin{aligned} \mathbb{E}[\mathbf{V}_{\beta_0}(\mathbf{X}_n)] &= \mathbb{E}[\mathbb{E}[\mathbf{V}_{\beta_0}(\mathbf{X}_n) | \mathcal{F}_{n-1}]] = \mathbb{E}[\mathbf{P}_{(\boldsymbol{\Sigma}_{n-1}, \boldsymbol{\beta}_{n-1})} \mathbf{V}_{\beta_0}(\mathbf{X}_{n-1})] \\ &\leq \lambda_{\epsilon, \beta_0} \mathbb{E}[\mathbf{V}_{\beta_0}(\mathbf{X}_{n-1})] + b_{\epsilon, \beta_0}. \end{aligned}$$

By iterating this majorisation, we get recursively

$$\begin{aligned} \mathbb{E}[\mathbf{V}_{\beta_0}(\mathbf{X}_n)] &\leq \lambda_{\epsilon, \beta_0}^n \mathbb{E}[\mathbf{V}_{\beta_0}(\mathbf{X}_0)] + b \sum_{k=1}^n \lambda_{\epsilon, \beta_0}^k \\ &\leq \lambda_{\epsilon, \beta_0} \mathbb{E}[\mathbf{V}_{\beta_0}(\mathbf{X}_0)] + \frac{b}{1 - \lambda_{\epsilon, \beta_0}}. \end{aligned}$$

Because the term on the right is independent from  $n$  and since, by assumption,  $\mathbb{E}\mathbf{V}_{\beta_0}(\mathbf{X}_0) < \infty$ , the proof is concluded.  $\square$

The following Lemma is adapted from [10, Lemma 4.2].

**Lemma 22.** Assume **(A1)**. Let  $\epsilon > 0$ ,  $\beta_0 > 0$ ,  $\eta > 0$ . For any  $(\boldsymbol{\Sigma}, \boldsymbol{\beta}) \in \mathcal{K}_{\epsilon, \beta_0, \eta} \stackrel{\text{def}}{=} \mathcal{M}_+^L(d, \epsilon) \times \mathcal{K}_{\beta_0, \eta}$ , let  $F_{(\boldsymbol{\Sigma}, \boldsymbol{\beta})} : \mathbf{X}^L \rightarrow \mathbb{R}^+$  be a measurable function such that

$$(56) \quad \sup_{(\boldsymbol{\Sigma}, \boldsymbol{\beta}) \in \mathcal{K}_{\epsilon, \beta_0, \eta}} \|F_{(\boldsymbol{\Sigma}, \boldsymbol{\beta})}\|_{\mathbf{V}_{\beta_0}} < +\infty.$$

Define  $\hat{F}_{(\boldsymbol{\Sigma}, \boldsymbol{\beta})}$  the solution of the Poisson equation (see Corollary 2)

$$\hat{F}_{(\boldsymbol{\Sigma}, \boldsymbol{\beta})} \stackrel{\text{def}}{=} \sum_{n \geq 0} \mathbf{P}_{(\boldsymbol{\Sigma}, \boldsymbol{\beta})}^n \{F_{(\boldsymbol{\Sigma}, \boldsymbol{\beta})} - \pi_{\boldsymbol{\beta}}(F_{(\boldsymbol{\Sigma}, \boldsymbol{\beta})})\}.$$

There exist a constant  $L_{\epsilon, \beta_0} < \infty$  such that, for any  $(\boldsymbol{\Sigma}, \boldsymbol{\beta}), (\boldsymbol{\Sigma}', \boldsymbol{\beta}') \in \mathcal{K}_{\epsilon, \beta_0, \eta}$ ,

$$(57) \quad \|\pi_{\boldsymbol{\beta}} - \pi_{\boldsymbol{\beta}'}\|_{\mathbf{V}_{\beta_0}} \leq L_{\epsilon, \beta_0} D_{\mathbf{V}_{\beta_0}} [(\boldsymbol{\Sigma}, \boldsymbol{\beta}), (\boldsymbol{\Sigma}', \boldsymbol{\beta}')],$$

and

$$\begin{aligned} (58) \quad &\|\mathbf{P}_{(\boldsymbol{\Sigma}, \boldsymbol{\beta})} \hat{F}_{(\boldsymbol{\Sigma}, \boldsymbol{\beta})} - \mathbf{P}_{(\boldsymbol{\Sigma}', \boldsymbol{\beta}')} \hat{F}_{(\boldsymbol{\Sigma}', \boldsymbol{\beta}')}\|_{\mathbf{V}_{\beta_0}} \\ &\leq L_{\epsilon, \beta_0} \left\{ \|F_{(\boldsymbol{\Sigma}, \boldsymbol{\beta})} - F_{(\boldsymbol{\Sigma}', \boldsymbol{\beta}')} \|_{\mathbf{V}_{\beta_0}} \right. \\ &\quad \left. + \sup_{(\boldsymbol{\Sigma}, \boldsymbol{\beta}) \in \mathcal{K}_{\epsilon, \beta_0, \eta}} \|F_{(\boldsymbol{\Sigma}, \boldsymbol{\beta})}\|_{\mathbf{V}_{\beta_0}} D_{\mathbf{V}_{\beta_0}} [(\boldsymbol{\Sigma}, \boldsymbol{\beta}), (\boldsymbol{\Sigma}', \boldsymbol{\beta}')] \right\}. \end{aligned}$$

We use repeatedly the following elementary result on a projection to a closed convex set.

**Lemma 23.** Let  $\mathbf{E}$  be an Euclidean space. Given any nonempty closed convex set  $\mathbf{K} \subset \mathbf{E}$ , denote by  $\Pi_{\mathbf{K}}$  the projection on the set  $\mathbf{K}$ . For any  $(x, y) \in \mathbf{E} \times \mathbf{E}$ ,  $\|\Pi_{\mathbf{K}}(x) - \Pi_{\mathbf{K}}(y)\| \leq \|x - y\|$ , where  $\|\cdot\|$  is the Euclidean norm.

**Lemma 24.** *Assume (A1) and  $\sup_{n \in \mathbb{N}} \gamma_{n,2} \leq 1$ , where  $\gamma_{n,2}$  is defined in (18) and (19). Then, for all  $\kappa > 0$  and  $\alpha \in [0, 1]$ , there exists constant  $K_{\kappa, \alpha, \epsilon, \beta_0, \eta} < \infty$  such that for any  $n \in \mathbb{N}$  it holds*

$$\begin{aligned} D_{\mathbf{V}_{\beta_0}^\alpha} [(\boldsymbol{\Sigma}_n, \boldsymbol{\beta}_n), (\boldsymbol{\Sigma}_{n+1}, \boldsymbol{\beta}_{n+1})] \\ \leq K_{\kappa, \epsilon, \beta_0, \eta} \gamma_{n+1} \left[ \mathbf{V}_{\beta_0}^\kappa(\mathbf{X}_{n+1}) + \sum_{k=0}^n a_{n,k} \mathbf{V}_{\beta_0}^\kappa(\mathbf{X}_k) \right], \end{aligned}$$

where  $\gamma_n = \sum_{i=1}^3 \gamma_{n,i}$ , and  $a_{n,n} \stackrel{\text{def}}{=} \gamma_{n,2}$  and for  $k \in \{0, \dots, n-1\}$

$$(59) \quad a_{n,k} \stackrel{\text{def}}{=} \gamma_{k,2} \prod_{j=k+1}^n (1 - \gamma_{j,2}).$$

*Proof.* According to Theorem 3, under (A1), there exists a constant  $K_{\epsilon, \alpha, \beta_0, \eta}$  such that

$$D_{\mathbf{V}_{\beta_0}^\alpha} [(\boldsymbol{\Sigma}_n, \boldsymbol{\beta}_n), (\boldsymbol{\Sigma}_{n-1}, \boldsymbol{\beta}_{n-1})] \leq K_{\epsilon, \alpha, \beta_0, \eta} \{ |\boldsymbol{\beta}_n - \boldsymbol{\beta}_{n+1}| + |\boldsymbol{\Sigma}_n - \boldsymbol{\Sigma}_{n+1}| \}.$$

For any  $\ell \in \{1, \dots, L-1\}$  consider  $|\rho_{n+1}^{(\ell)} - \rho_n^{(\ell)}|$ , where  $\{\rho_n^{(\ell)}\}$  is defined by (14). Since, by (15),  $|H^{(\ell)}(\rho^{(1:\ell)}, x)| \leq 1$  for any  $\rho^{(1:\ell)} \in \mathbb{R}^\ell$  and  $x \in \mathbf{X}$ , applying Lemma 23 we obtain

$$(60) \quad |\rho_{n+1}^{(\ell)} - \rho_n^{(\ell)}| \leq \gamma_{n,1} |H^{(\ell)}(\rho_{n-1}^{(1:\ell)}, \mathbf{X}_n)| \leq \gamma_{n,1}.$$

Define the function

$$\boldsymbol{\beta} : \rho^{(1:L-1)} \rightarrow (1, \beta^{(2)}(\rho^{(1)}), \dots, \beta^{(L)}(\rho^{(1:L-1)}))$$

where the functions  $\beta^{(\ell)}$  are defined in (13). The function  $\boldsymbol{\beta}$  is continuously differentiable. By definition (14), for all  $n \in \mathbb{N}$ ,  $\boldsymbol{\rho}_n \in [\underline{\rho}, \bar{\rho}]^{L-1}$ . Hence (60) implies that there exists  $K < \infty$  such that

$$(61) \quad |\boldsymbol{\beta}_n - \boldsymbol{\beta}_{n+1}| = |\boldsymbol{\beta}(\boldsymbol{\rho}_n) - \boldsymbol{\beta}(\boldsymbol{\rho}_{n+1})| \leq K \gamma_{n,1}.$$

Now consider  $|\boldsymbol{\Sigma}_n - \boldsymbol{\Sigma}_{n+1}|$ . By definition (21) we get

$$|\boldsymbol{\Sigma}_n^{(\ell)} - \boldsymbol{\Sigma}_{n+1}^{(\ell)}| \leq \exp(T_n^{(\ell)}) |\Gamma_n^{(\ell)} - \Gamma_{n+1}^{(\ell)}| + |\exp(T_n^{(\ell)}) - \exp(T_{n+1}^{(\ell)})| |\Gamma_{n+1}^{(\ell)}|.$$

The scale adaptation procedure (20) by Lemma 23 satisfies

$$|T_n^{(\ell)} - T_{n+1}^{(\ell)}| \leq \gamma_{n+1,3},$$

which implies that there exists constant  $K < \infty$  such that  $|\exp(T_n^{(\ell)}) - \exp(T_{n+1}^{(\ell)})| \leq K \gamma_{n+1,3}$ . By (18) and Lemma 23 we obtain

$$\begin{aligned} |\Gamma_n^{(\ell)} - \Gamma_{n+1}^{(\ell)}| &\leq \gamma_{n+1,2} \left| (X_{n+1}^{(\ell)} - \mu_n^{(\ell)})(X_{n+1}^{(\ell)} - \mu_n^{(\ell)})^T + \Gamma_n^{(\ell)} \right| \\ &\leq \gamma_{n+1,2} \left[ 2|X_{n+1}^{(\ell)}|^2 + 2|\mu_n^{(\ell)}|^2 + |\Gamma_n^{(\ell)}| \right]. \end{aligned}$$



By definition (18) and (20)  $\exp(T_n)$  and  $|\Gamma_n|$  are uniformly bounded for all  $n \in \mathbb{N}$ . Hence, gathering all terms, there exists a constant  $K < \infty$  such that

$$D\mathbf{V}_{\beta_0}^\alpha [(\boldsymbol{\Sigma}_n, \boldsymbol{\beta}_n), (\boldsymbol{\Sigma}_{n-1}, \boldsymbol{\beta}_{n-1})] \leq K\gamma_{n+1} \left[ |X_{n+1}^{(\ell)}|^2 + |\mu_n^{(\ell)}|^2 \right],$$

where  $\gamma_n = \sum_{i=1}^3 \gamma_{n,i}$ . Under assumption  $\sup_{n \in \mathbb{N}} \gamma_{n,2} \leq 1$ , by (19), we get that for any  $n \in \mathbb{N}$   $\mu_n^{(\ell)} = \sum_{k=0}^n a_{n,k} X_{n-k}^{(\ell)}$ , where the positive weights  $a_{n,k}$ ,  $k \in \{0, \dots, n\}$  are defined in (59). Because  $\sum_{k=0}^n a_{n,k} = 1$ , the Jensen inequality implies

$$|\mu_n^{(\ell)}|^2 \leq \left( \sum_{k=0}^n a_{n,k} |X_{n-k}^{(\ell)}| \right)^2 \leq \sum_{k=0}^n a_{n,k} |X_{n-k}^{(\ell)}|^2.$$

Finally, under **(A1)** for any  $\kappa > 0$  there exists  $K_\kappa$  such that, for all  $\mathbf{x} \in \mathbf{X}^L$ ,  $|\mathbf{x}|^2 \leq K_\kappa \mathbf{V}_{\beta_0}^\kappa(\mathbf{x})$ .  $\square$

**Lemma 25.** *Assume **(A1)** and  $\sup_{n \in \mathbb{N}} \gamma_{n,2} \leq 1$ . For any non-negative sequence  $\{b_n\}_{n \geq 0}$  satisfying  $\sum_{n \geq 1} b_n(\gamma_{n,1} + \gamma_{n,2} + \gamma_{n,3}) < \infty$  and for all  $\alpha \in (0, 1)$ , we have*

$$\sum_{n=1}^{\infty} b_n D\mathbf{V}_{\beta_0}^\alpha [(\boldsymbol{\Sigma}_n, \boldsymbol{\beta}_n), (\boldsymbol{\Sigma}_{n-1}, \boldsymbol{\beta}_{n-1})] \mathbf{V}_{\beta_0}^\alpha(\mathbf{X}_n) < +\infty \text{ a.s. .}$$

*Proof.* Since  $\sup_{n \in \mathbb{N}} \gamma_{n,2} \leq 1$ , under **(A1)** by Lemma 24 for all  $\kappa > 0$  there exists  $K < \infty$  such that

$$\begin{aligned} D\mathbf{V}_{\beta_0}^\alpha [(\boldsymbol{\Sigma}_n, \boldsymbol{\beta}_n), (\boldsymbol{\Sigma}_{n-1}, \boldsymbol{\beta}_{n-1})] \\ \leq K\gamma_{n+1} \left[ \mathbf{V}_{\beta_0}^\kappa(\mathbf{X}_{n+1}) + \sum_{k=0}^n a_{n,k} \mathbf{V}_{\beta_0}^\kappa(\mathbf{X}_k) \right], \end{aligned}$$

where  $\gamma_n = \sum_{i=1}^3 \gamma_{n,i}$  and the triangular array  $\{a_{n,k}\}_{k=0}^n$  is defined in (59). Set  $\kappa = 1 - \alpha$ . Since  $\sum_{n \geq 1} b_n \gamma_n < \infty$ , it is enough to show that

$$A \stackrel{\text{def}}{=} \sup_{n \in \mathbb{N}} \mathbb{E} \left\{ \left( \mathbf{V}_{\beta_0}^{1-\alpha}(\mathbf{X}_{n+1}) + \sum_{k=0}^n a_{n,k} \mathbf{V}_{\beta_0}^{1-\alpha}(\mathbf{X}_k) \right) \mathbf{V}_{\beta_0}^\alpha(\mathbf{X}_{n+1}) \right\} < \infty.$$

By Hölder inequality we get for all  $k, n \in \mathbb{N}$

$$\mathbb{E} \{ \mathbf{V}_{\beta_0}^{1-\alpha}(\mathbf{X}_k) \mathbf{V}_{\beta_0}^\alpha(\mathbf{X}_{n+1}) \} \leq \{ \mathbb{E} \mathbf{V}_{\beta_0}(\mathbf{X}_{n+1}) \}^\alpha \{ \mathbb{E} \mathbf{V}_{\beta_0}(\mathbf{X}_k) \}^{1-\alpha}.$$

Since the weights  $a_{n,k}$  are non-negative and  $\sum_{k=0}^n a_{n,k} = 1$ , we get  $A \leq 2 \sup_{n \in \mathbb{N}} \mathbb{E} \mathbf{V}_{\beta_0}(\mathbf{X}_n)$  the proof is concluded by applying Lemma 21.  $\square$

*Proof of Theorem 5.* According to [10, Corollary 2.9], we need to check that **(i)**  $\sup_{n \geq 0} \mathbb{E}[\mathbf{V}(\mathbf{X}_n)] < +\infty$ , and **(ii)** there exists  $\alpha \in (0, 1)$  such that

$$\sum_{n=1}^{\infty} n^{-1} D\mathbf{V}_{\beta_0}^\alpha [(\boldsymbol{\Sigma}_n, \boldsymbol{\beta}_n), (\boldsymbol{\Sigma}_{n-1}, \boldsymbol{\beta}_{n-1})] \mathbf{V}_{\beta_0}^\alpha(\mathbf{X}_n) < +\infty \text{ a.s. ,}$$

which follow from Lemma 21 and Lemma 25, respectively.  $\square$

**5.4. Proof of Proposition 7.** In order to prove the existence and uniqueness of the root of the mean field  $\mathbf{h}$ , we first introduce some notation and prove the key Lemma 26.

Below, we omit the set  $\mathbf{X}$  from the integrals. Let us define a symmetric function  $\tilde{h} : (0, 1]^2 \rightarrow [0, 1]$  as follows

$$(62) \quad \tilde{h}(u, v) = \iint \left( 1 \wedge \frac{\pi^u(x)\pi^v(y)}{\pi^u(y)\pi^v(x)} \right) \frac{\pi^v(dx)}{Z(v)} \frac{\pi^u(dy)}{Z(u)}.$$

Note that for all  $(u, v) \in (0, 1]^2$ , we get

$$(63) \quad \tilde{h}(u, v) = \int f_v(\pi(y)) \frac{\pi^u(y)}{Z(u)} dy = \int f_u(\pi(y)) \frac{\pi^v(y)}{Z(v)} dy,$$

where for  $v \in (0, 1]$ ,

$$(64) \quad f_v(z) \stackrel{\text{def}}{=} \int \frac{\pi^v(dx)}{Z(v)} [\mathbb{1}_{\{\pi \leq z\}}(x) + \mathbb{1}_{\{\pi < z\}}(x)].$$

**Lemma 26.** *Assume that  $\lambda^{\text{Leb}}(\mathbf{X}) = \infty$  and the density  $\pi$  is positive, bounded, and  $\int \pi^\kappa(x) \lambda^{\text{Leb}}(dx) < \infty$  for all  $0 < \kappa \leq 1$ . Then, for all  $v \in (0, 1]$  the function  $u \mapsto \tilde{h}(u, v)$  restricted to  $(0, v)$  is differentiable and monotonic with the following derivative*

$$\begin{aligned} \frac{\partial \tilde{h}}{\partial u}(u, v) &= \frac{1}{2} \iint [f_v(\pi(y)) - f_v(\pi(z))] \\ &\quad \times [\log(\pi(y)) - \log(\pi(z))] \frac{\pi^u(dy)}{Z(u)} \frac{\pi^u(dz)}{Z(u)}. \end{aligned}$$

Moreover,  $\lim_{u \rightarrow 0} \tilde{h}(u, v) = 0$  and  $\lim_{u \rightarrow v} \tilde{h}(u, v) = 1$ .

*Proof.* We will first show that, for any bounded measurable function  $f$  and  $u \in (0, 1)$ ,

$$(65) \quad \frac{d}{du} \left( \int_{\mathbf{X}} f(y) \pi^u(dy) \right) = \int_{\mathbf{X}} f(y) \log(\pi(y)) \pi^u(dy).$$

Since  $\pi$  is bounded, there exists a function  $J : \mathbf{X} \rightarrow [0, \infty)$  such that  $\pi(x) = c \exp(-J(x))$ . By the dominated convergence theorem, we only need to show that  $|h^{-1}(\pi^{u+h}(y) - \pi^u(y))|$  is bounded uniformly for  $h$  in some neighbourhood of 0 by an integrable function depending only on  $y$ . We may write

$$\left| \frac{\pi^{u+h}(y) - \pi^u(y)}{h} \right| = c \exp(-uJ(y)) \frac{|\exp(-hJ(y)) - 1|}{|h|}.$$

Applying the mean value theorem we obtain  $|\exp(-hJ(y)) - 1| \leq |h| \exp(|h|J(y))$ . Hence for all  $|h| \leq \frac{u}{2}$  we get that

$$\left| \frac{\pi^{u+h}(y) - \pi^u(y)}{h} \right| \leq c \exp \left( -\frac{uJ(y)}{2} \right) = c \pi^{u/2}(y),$$

which concludes the proof of (65).

For any given  $v$ , using (65), we compute the partial derivative of  $\tilde{h}$  with respect to  $u$

$$\begin{aligned} \frac{\partial \tilde{h}}{\partial u}(u, v) &= \int f_v(\pi(y)) \log(\pi(y)) \frac{\pi^u(dy)}{Z(u)} - \int f_v(\pi(y)) \frac{\pi^u(dy)}{Z(u)} \int \log(\pi(y)) \frac{\pi^u(dy)}{Z(u)} \\ &= \frac{1}{2} \int [f_v(\pi(y)) - f_v(\pi(z))] [\log(\pi(y)) - \log(\pi(z))] \frac{\pi^u(dy)}{Z(u)} \frac{\pi^u(dz)}{Z(u)}. \end{aligned}$$

Since the functions  $z \mapsto f_v(z)$  and  $z \mapsto \log(z)$  are non-decreasing,  $[f_v(\pi(y)) - f_v(\pi(z))] [\log(\pi(y)) - \log(\pi(z))] \geq 0$  for all  $y, z$ . Moreover, because  $\pi$  is positive the  $(\pi \times \pi)$ -measure of the sets  $\{y, z \in \mathbf{X} : \pi(y) \neq \pi(z)\}$  must be positive due to  $\lambda^{\text{Leb}}(\mathbf{X}) = \infty$ . Hence  $\frac{\partial \tilde{h}}{\partial u}(u, v) > 0$  for  $u \in (0, v)$  and this completes the proof of the first part.

Since  $\pi^u \vee \pi^v$  is integrable, the dominated convergence theorem implies  $\tilde{h}(u, v) \rightarrow 1$  as  $u \rightarrow v$ . For the second limit consider (63). For any  $\varepsilon > 0$  we can find a  $\delta > 0$  such that  $f_v(\delta) \leq \varepsilon/2$ . Therefore,

$$\int f_v(\pi(y)) \frac{\pi^u(dy)}{Z(u)} \leq \frac{\varepsilon}{2} \int \frac{\pi^u(dy)}{Z(u)} + \frac{\int \mathbb{1}_{\{\pi > \delta\}}(y) \pi^u(dy)}{\int \pi^u(dy)}.$$

There exists a constant  $c < \infty$  such that  $\int \mathbb{1}_{\{\pi > \delta\}}(y) \pi^u(dy) \leq c$  for all  $u \in [0, 1]$  and

$$\int \pi^u(dy) \geq \int \pi^u(dy) \mathbb{1}_{\{\pi \leq 1\}}(y).$$

Observe that, for any  $y \in \mathbf{X}$ , the function  $u \mapsto \pi^u(y) \mathbb{1}_{\{\pi \leq 1\}}(y)$  is non-increasing. Therefore, using the monotone convergence theorem and  $\lambda^{\text{Leb}}(\mathbf{X}) = \infty$ , we get  $\lim_{u \rightarrow 0} \int \pi^u(dy) = \infty$ . Hence we can find  $u_0$  such that for all  $u < u_0$  the normalising constant  $\int \pi^u(dy) \geq c/\varepsilon$  and therefore  $\int f_v(\pi(y)) \frac{\pi^u(dy)}{Z(u)} \leq \varepsilon$ .  $\square$

*Proof of Proposition 7.* We recall that  $h^{(1)}(\rho)$  depends only on  $\rho^{(1)}$  and we may write  $h^{(1)}(\rho) = \tilde{h}(\exp(-\exp(\rho^{(1)})), 1) - \alpha^*$ , where  $\tilde{h}$  is defined in (62). By Lemma 26 the function  $u \mapsto \tilde{h}(u, 1)$  is strictly monotonic on  $[0, 1]$  and  $\lim_{u \rightarrow 0} \tilde{h}(u, 1) = 0$  and  $\lim_{u \rightarrow 1} \tilde{h}(u, 1) = 1$ . Since  $u \rightarrow \tilde{h}(u, 1)$  is continuous, there exists a unique  $\hat{\rho}^{(1)}$  such that  $h^{(1)}(\hat{\rho}^{(1)}, \rho^{(2)}, \dots, \rho^{(L-1)}) = h^{(1)}(\hat{\rho}^{(1)}) = 0$ . We proceed by induction and assume the existence of unique  $\hat{\rho}^{(1)}, \dots, \hat{\rho}^{(\ell-1)}$  such that for any  $k < \ell$  we have  $h^{(k)}(\hat{\rho}^{(1)}, \dots, \hat{\rho}^{(k)}) = 0$ . Denoting

$$(66) \quad u(\rho) = \exp(-\exp(\rho))$$

$$(67) \quad v_{\ell-1}(\rho^{(1:\ell-1)}) = \exp\left(-\sum_{i=1}^{\ell-1} \exp(\rho^{(i)})\right),$$

we may write

$$h^{(\ell)}(\hat{\rho}^{(1:\ell-1)}, \rho^{(\ell)}) = \tilde{h}(u(\rho^{(\ell)})v_{\ell-1}(\hat{\rho}^{(1:\ell-1)}), v_{\ell-1}(\hat{\rho}^{(1:\ell-1)})) - \alpha^*.$$

We conclude as before from the monotonicity of  $u \mapsto \tilde{h}(u, v)$  and the limits  $\lim_{u \rightarrow 0} \tilde{h}(u, v) = 0$  and  $\lim_{u \rightarrow v} \tilde{h}(u, v) = 1$  (see Lemma 26).  $\square$

**5.5. Proof of Theorem 10.** Our proof of convergence of the temperature adaptation employs classical convergence results on stochastic approximation. It is, however, not easy to construct a ‘global’ Lyapunov function for the mean field  $\mathbf{h}$ , but the specific structure of the problem allows to deduce the convergence recursively for  $\rho^{(1)}, \dots, \rho^{(L-1)}$ .

In this section, for notational simplicity,  $\mathbf{P}_{(\boldsymbol{\Sigma}, \beta(\rho))} = \mathbf{P}_{(\boldsymbol{\Sigma}, \rho)}$  and  $\pi_{\beta(\rho)} = \pi_{\rho}$ . For any  $\ell = 1, \dots, L-1$  we rewrite (14) as follows

$$(68) \quad \rho_n^{(\ell)} = \Pi_{\rho}[\rho_{n-1}^{(\ell)} + \gamma_{n,1}g^{(\ell)}(\rho_{n-1}^{(\ell)}) + \gamma_{n,1}\varepsilon_n^{(\ell)} + \gamma_{n,1}r_n^{(\ell)}],$$

where  $\Pi_{\rho}$  is defined in (14) and

$$(69) \quad g^{(\ell)}(\rho) = \tilde{h}(u(\rho)v_{\ell-1}(\hat{\rho}^{(1:\ell-1)}), v_{\ell-1}(\hat{\rho}^{(1:\ell-1)})) - \alpha^*$$

$$(70) \quad \varepsilon_n^{(\ell)} = H_{\rho_{n-1}^{(\ell)}}^{(\ell)}(\mathbf{X}_n) - \pi_{\rho_{n-1}^{(\ell)}}[H_{\rho_{n-1}^{(\ell)}}^{(\ell)}(\mathbf{X}_n)]$$

$$(71) \quad r_n^{(\ell)} = \tilde{h}(u(\rho_{n-1}^{(\ell)})v_{\ell-1}(\rho_n^{(1:\ell-1)}), v_{\ell-1}(\rho_n^{(1:\ell-1)})) \\ - \tilde{h}(u(\rho_{n-1}^{(\ell)})v_{\ell-1}(\hat{\rho}^{(1:\ell-1)}), v_{\ell-1}(\hat{\rho}^{(1:\ell-1)})),$$

where  $\tilde{h}$  is defined in (62),  $H_{\rho^{(1:\ell-1)}}^{(\ell)}(\mathbf{x})$  is a shorthand notation for  $H^{(\ell)}(\rho^{(1:\ell-1)}, \mathbf{x})$  defined in (15),  $u$  and  $v_{\ell-1}$  are defined in (66) and (67), respectively, and, by convention,  $v_0 = 1$ .

We decompose the term  $\varepsilon_n^{(\ell)}$  as the sum of a martingale difference term and a remainder term that goes to zero. To do this, we use the Poisson decomposition. For  $(\boldsymbol{\Sigma}, \rho) \in \mathcal{M}_+^L(d, \varepsilon) \times [\underline{\rho}, \bar{\rho}]^{L-1}$  defined in Section 2, denote by  $\hat{H}_{(\boldsymbol{\Sigma}, \rho)}^{(\ell)}$  the solution of the Poisson equation

$$\hat{H}_{(\boldsymbol{\Sigma}, \rho)}^{(\ell)}(\mathbf{x}) - \mathbf{P}_{(\boldsymbol{\Sigma}, \rho)}\hat{H}_{(\boldsymbol{\Sigma}, \rho)}^{(\ell)}(\mathbf{x}) = H_{\rho^{(1:\ell)}}^{(\ell)}(\mathbf{x}) - \pi_{\rho}[H_{\rho^{(1:\ell)}}^{(\ell)}],$$

which exists by Corollary 2. Hence,  $\varepsilon_n^{(\ell)} = \delta M_n^{(\ell)} + \kappa_n^{(\ell)}$  where

$$\delta M_n^{(\ell)} \stackrel{\text{def}}{=} \hat{H}_{\boldsymbol{\Sigma}_{n-1}, \rho_{n-1}}^{(\ell)}(\mathbf{X}_n) - \mathbf{P}_{(\boldsymbol{\Sigma}_{n-1}, \rho_{n-1})}\hat{H}_{\boldsymbol{\Sigma}_{n-1}, \rho_{n-1}}^{(\ell)}(\mathbf{X}_{n-1}) \\ \kappa_n^{(\ell)} \stackrel{\text{def}}{=} \mathbf{P}_{(\boldsymbol{\Sigma}_{n-1}, \rho_{n-1})}\hat{H}_{\boldsymbol{\Sigma}_{n-1}, \rho_{n-1}}^{(\ell)}(\mathbf{X}_{n-1}) - \mathbf{P}_{(\boldsymbol{\Sigma}_{n-1}, \rho_{n-1})}\hat{H}_{\boldsymbol{\Sigma}_{n-1}, \rho_{n-1}}^{(\ell)}(\mathbf{X}_n).$$

**Lemma 27.** Assume **(A1)** - **(A3)** and  $\mathbb{E}\mathbf{V}_{\beta_0}(\mathbf{X}_0) < \infty$ . For all  $\ell \in \{1, \dots, L-1\}$  and  $T < \infty$ , it holds

$$(72) \quad \lim_{n \rightarrow \infty} \sup_{n \leq k \leq m(n, T)} \left| \sum_{i=n}^k \gamma_{i,1} \delta M_i^{(\ell)} \right| = 0 \quad a.s.,$$

$$(73) \quad \lim_{n \rightarrow \infty} \sup_{n \leq k \leq m(n, T)} \left| \sum_{i=n}^k \gamma_{i,1} \kappa_i^{(\ell)} \right| = 0 \quad a.s.,$$

where

$$m(n, T) \stackrel{\text{def}}{=} \max \left\{ j > n : \sum_{i=n+1}^j \gamma_{i,1} \leq T \right\}.$$

*Proof.* Consider (72). Since  $\delta M_i^{(\ell)}$  are martingale increments, Doob's inequality implies

$$(74) \quad \mathbb{E} \left( \sup_{k \geq n} \left| \sum_{i=n}^k \gamma_{i,1} \delta M_i^{(\ell)} \right| \right)^2 \leq K \sum_{i=n}^{\infty} \gamma_{i,1}^2 \mathbb{E} \left[ \left| \hat{H}_{\Sigma_{i-1}, \rho_{i-1}}^{(\ell)}(\mathbf{X}_i) - \mathbf{P}_{(\Sigma_{i-1}, \rho_{i-1})} \hat{H}_{\Sigma_{i-1}, \rho_{i-1}}^{(\ell)}(\mathbf{X}_{i-1}) \right|^2 \right].$$

Since  $|H^{(\ell)}| \leq 1$ , Corollary 2 yields

$$(75) \quad \sup_{(\Sigma, \rho) \in \mathcal{M}_+^L(d, \epsilon) \times [\underline{\rho}, \bar{\rho}]^{L-1}} |\hat{H}_{(\Sigma, \rho)}^{(\ell)}(\mathbf{x})| \leq K \mathbf{V}_{\beta_0}^{1/2}(\mathbf{x})$$

$$(76) \quad \sup_{(\Sigma, \rho) \in \mathcal{M}_+^L(d, \epsilon) \times [\underline{\rho}, \bar{\rho}]^{L-1}} |\mathbf{P}_{(\Sigma, \rho)} \hat{H}_{(\Sigma, \rho)}^{(\ell)}(\mathbf{x})| \leq K \mathbf{V}_{\beta_0}^{1/2}(\mathbf{x}).$$

Hence by (74) we obtain

$$\mathbb{E} \left( \sup_{k \geq n} \left| \sum_{i=n}^k \gamma_{i,1} \delta M_i^{(\ell)} \right| \right)^2 \leq K \sum_{i=n}^{\infty} \gamma_{i,1}^2 \mathbb{E} \mathbf{V}_{\beta_0}(\mathbf{X}_i).$$

Lemma 21 shows that  $\sup_{i \geq 0} \mathbb{E} \mathbf{V}_{\beta_0}(\mathbf{X}_i) < \infty$ , and the proof of (72) is concluded under the step size condition **(A3)**.

Now consider (73). Decompose  $\sum_{i=n}^k \gamma_{i,1} \kappa_i^{(\ell)} = R_{n,k}^{(\ell,1)} + R_{n,k}^{(\ell,2)} + R_{n,k}^{(\ell,3)}$  with

$$\begin{aligned} R_{n,k}^{(\ell,1)} &\stackrel{\text{def}}{=} \sum_{i=n-1}^{k-1} \gamma_{i+1,1} \left[ \mathbf{P}_{(\Sigma_i, \rho_i)} \hat{H}_{\Sigma_i, \rho_i}^{(\ell)}(\mathbf{X}_i) - \mathbf{P}_{(\Sigma_{i-1}, \rho_{i-1})} \hat{H}_{\Sigma_{i-1}, \rho_{i-1}}^{(\ell)}(\mathbf{X}_i) \right] \\ R_{n,k}^{(\ell,2)} &\stackrel{\text{def}}{=} \gamma_{n-1,1} \mathbf{P}_{(\Sigma_{n-2}, \rho_{n-2})} \hat{H}_{\Sigma_{n-2}, \rho_{n-2}}^{(\ell)}(\mathbf{X}_{n-1}) - \gamma_{k,1} \mathbf{P}_{(\Sigma_{k-1}, \rho_{k-1})} \hat{H}_{\Sigma_{k-1}, \rho_{k-1}}^{(\ell)}(\mathbf{X}_k) \\ R_{n,k}^{(\ell,3)} &\stackrel{\text{def}}{=} \sum_{i=n-1}^{k-1} (\gamma_{i+1,1} - \gamma_{i,1}) \mathbf{P}_{(\Sigma_{i-1}, \rho_{i-1})} \hat{H}_{\Sigma_{i-1}, \rho_{i-1}}^{(\ell)}(\mathbf{X}_i). \end{aligned}$$

By Lemma 22 we get  $|R_{n,k}^{(\ell,1)}| \leq R_n^{(\ell,1,1)} + R_n^{(\ell,1,2)}$  where

$$R_n^{(\ell,1,1)} \stackrel{\text{def}}{=} K \sum_{i=n-1}^{\infty} \gamma_{i+1,1} \mathbf{V}_{\beta_0}^{1/2}(\mathbf{X}_i) \|H_{\rho_i^{(1:\ell)}}^{(\ell)} - H_{\rho_{i-1}^{(1:\ell)}}^{(\ell)}\|_{\mathbf{V}_{\beta_0}^{1/2}}$$

$$R_n^{(\ell,1,2)} \stackrel{\text{def}}{=} \sum_{i=n-1}^{\infty} \gamma_{i+1,1} \mathbf{V}_{\beta_0}^{1/2}(\mathbf{X}_i) D_{\mathbf{V}_{\beta_0}^{1/2}} [(\boldsymbol{\Sigma}_i, \boldsymbol{\rho}_i), (\boldsymbol{\Sigma}_{i-1}, \boldsymbol{\rho}_{i-1})] .$$

Since  $H^{(\ell)}$  is locally Lipschitz with respect to  $\boldsymbol{\rho}$ , by Lemma 23 we obtain

$$\|H_{\rho_i^{(1:\ell)}}^{(\ell)} - H_{\rho_{i-1}^{(1:\ell)}}^{(\ell)}\|_{\mathbf{V}_{\beta_0}^{1/2}} \leq K \gamma_{i,1} .$$

Lemma 21 and **(A3)** imply  $\lim_{n \rightarrow \infty} R_n^{(\ell,1,1)} = 0$  a.s. and Lemma 25 shows that  $\lim_{n \rightarrow \infty} R_n^{(\ell,1,2)} = 0$  a.s.. By (76), Lemma 21 and **(A3)**, the sum  $\sum_k (\gamma_{k,1} \mathbf{P}_{(\boldsymbol{\Sigma}_{k-1}, \boldsymbol{\rho}_{k-1})} \hat{H}_{\boldsymbol{\Sigma}_{k-1}, \boldsymbol{\rho}_{k-1}}^{(\ell)}(\mathbf{X}_k))^2 < \infty$  a.s., implying that  $\sup_{k \geq n} |R_{n,k}^{(\ell,2)}| \xrightarrow{n \rightarrow \infty} 0$  a.s.. Finally, for the term  $R_{n,k}^3$ , (75), Lemma 21 and **(A3)** conclude the proof.  $\square$

*Proof of Theorem 10.* For any  $\ell = 1, \dots, L-1$ , define the local Lyapunov function  $w^{(\ell)} : \mathbb{R} \rightarrow \mathbb{R}$  through

$$(77) \quad w^{(\ell)}(\rho) := (\rho - \hat{\rho}^{(\ell)})^2 ,$$

where  $\hat{\rho}^{(\ell)}$  is the unique solution of the equation  $g^{(\ell)}(\rho) = 0$  where  $g^{(\ell)}$  is defined in (69). Observe that, by Lemma 26 and Proposition 7, for any  $\ell = 1, \dots, L-1$  the functions  $w^{(\ell)}$  satisfy, for all  $\rho \in \mathbb{R}$

$$\frac{\partial w^{(\ell)}}{\partial \rho}(\rho) \cdot g^{(\ell)}(\rho) \leq 0 ,$$

with equality if and only if  $\rho = \hat{\rho}^{(\ell)}$ . The projection set  $[\underline{\rho}, \bar{\rho}]$  obviously satisfies [18, A4.3.1] and since for all  $\ell = 1, \dots, L-1$ ,  $\underline{\rho} < \hat{\rho}^{(\ell)} < \bar{\rho}$ , the convergence cannot occur on the boundary of constraint set. Hence we only need to check the assumptions of [18, Theorem 6.1.1]. Since the mean field  $g^{(\ell)}$  is bounded and continuous, and depends only on  $\rho$  the conditions [18, A6.1.2, A6.1.6, A6.1.7] are satisfied. Condition [18, A6.1.1] is implied by the boundedness of  $H^{(\ell)}$ . Condition [18, A6.1.3] is implied by Lemma 27.

Once we show that the remainder term  $r_n^{(\ell)}$ , defined in (71) converges a.s. to zero as  $n \rightarrow \infty$ , Lemma 27 implies the condition [18, A6.1.4]. We will prove inductively that

$$(78) \quad \lim_{n \rightarrow \infty} |r_n^{(\ell)}| = 0 \quad a.s. .$$

Consider first the case  $\ell = 1$ . By (68) we get that  $r_n^{(1)} = 0$ . Assume that for  $\ell < k$  (78) holds, then by [18, Theorem 6.1.1] for any  $\ell < k$  we know that  $\rho_n^{(\ell)} \rightarrow \hat{\rho}^{(\ell)}$ . To simplify notation we denote  $u_n = u(\rho_n^{(k)})$ ,

$v_n = v^{(k)}(\rho_n^{(1:k-1)})$  and  $\hat{v} = v^{(k)}(\hat{\rho}^{(1:k-1)})$ . By (64), we get that  $f_u(z) \leq 2$  for all  $z$ . Hence, using (63), we get

$$\begin{aligned} |r_n^{(k)}| &\leq |\tilde{h}(u_n v_n, v_n) - \tilde{h}(u_n \hat{v}, v_n)| + |\tilde{h}(u_n \hat{v}, v_n) - \tilde{h}(u_n \hat{v}, \hat{v})| \\ &\leq \left| \int f_{v_n}(\pi(y)) \left[ \frac{\pi^{u_n v_n}(y)}{Z(u_n v_n)} - \frac{\pi^{u_n \hat{v}}(y)}{Z(u_n \hat{v})} \right] dy \right| \\ &\quad + \left| \int f_{u_n \hat{v}}(\pi(y)) \left[ \frac{\pi^{v_n}(y)}{Z(v_n)} - \frac{\pi^{\hat{v}}(y)}{Z(\hat{v})} \right] dy \right| \\ &\leq 2 \int \left| \frac{\pi^{u_n v_n}(y)}{Z(u_n v_n)} - \frac{\pi^{u_n \hat{v}}(y)}{Z(u_n \hat{v})} \right| dy + 2 \int \left| \frac{\pi^{v_n}(y)}{Z(v_n)} - \frac{\pi^{\hat{v}}(y)}{Z(\hat{v})} \right| dy \\ &= 2 \left\{ \left\| \frac{\pi^{u_n v_n}}{Z(u_n v_n)} - \frac{\pi^{u_n \hat{v}}}{Z(u_n \hat{v})} \right\|_{\text{TV}} + \left\| \frac{\pi^{v_n}}{Z(v_n)} - \frac{\pi^{\hat{v}}}{Z(\hat{v})} \right\|_{\text{TV}} \right\}. \end{aligned}$$

By [9, Lemma A.1] we get that

$$|r_n^{(k)}| \leq 4 \{ |\log(Z(u_n v_n)) - \log(Z(u_n \hat{v}))| + |\log(Z(v_n)) - \log(Z(\hat{v}))| \},$$

where the normalising constants  $Z$  are defined in (2). Since  $\alpha \mapsto \log(Z(\alpha))$  is locally Lipschitz,  $u_n$  and  $v_n$  are in compact set we get that exists  $K < \infty$  such that

$$|r_n^{(k)}| \leq K(1 + u_n)|v_n - \hat{v}|.$$

By (14) and (66) we obtain that  $\sup_n |u_n| < \infty$ . Hence, since  $\rho^{(1:\ell-1)} \mapsto v^{(k)}(\rho^{(1:\ell-1)})$  is continuous, by induction assumption, we conclude that  $|r_n^{(k)}| \rightarrow 0$  a.s..  $\square$

#### ACKNOWLEDGEMENTS

The first two authors were supported by the French National Research Agency under the contract ANR-08-BLAN-0218 “BigMC”. The third author was supported by the Academy of Finland (project 250575) and by the Finnish Academy of Science and Letters, Vilho, Yrjö and Kalle Väisälä Foundation.

#### REFERENCES

- [1] C. Andrieu and É. Moulines. On the ergodicity properties of some adaptive MCMC algorithms. *Ann. Appl. Probab.*, 16(3):1462–1505, 2006.
- [2] C. Andrieu and J. Thoms. A tutorial on adaptive MCMC. *Statist. Comput.*, 18(4):343–373, 2008.
- [3] Y. Atchadé and G. Fort. Limit theorems for some adaptive MCMC algorithms with subgeometric kernels. *Bernoulli*, 16(1):116–154, 2010.
- [4] Y. F. Atchadé and J. S. Rosenthal. On adaptive Markov chain Monte Carlo algorithms. *Bernoulli*, 11(5):815–828, 2005.

- [5] Y. F. Atchadé, G. O. Roberts, and J. S. Rosenthal. Towards optimal scaling of Metropolis-coupled Markov chain Monte Carlo. *Statist. Comput.*, 21(4):555–568, 2011.
- [6] J. D. Banfield and A. E. Raftery. Ice floe identification in satellite images using mathematical morphology and clustering about principal curves. *J. Amer. Statist. Assoc.*, 87:7–16, 1992.
- [7] M. Baragatti, A. Grimaud, and D. Pommeret. Parallel tempering with equi-energy moves. *Statist. Comput.*, to appear.
- [8] L. Bornn, P. E. Jacob, P. D. Moral, and A. Doucet. An adaptive interacting Wang-Landau algorithm for automatic density exploration. Technical report, 2011. URL <http://arxiv.org/abs/1109.3829v2>.
- [9] R. Douc, E. Moulines, and J. S. Rosenthal. Quantitative bounds on convergence of time-inhomogeneous Markov chains. *Ann. Appl. Prob.*, 2004:1643–1665, 2002.
- [10] G. Fort, E. Moulines, and P. Priouret. Convergence of adaptive and interacting Markov chain Monte Carlo algorithms. *Ann. Statist.*, 39(6):3262–3289, 2011.
- [11] W. R. Gilks, S. Richardson, and D. J. Spiegelhalter. *Markov Chain Monte Carlo in Practice*. Chapman & Hall/CRC, Boca Raton, Florida, 1998. ISBN 0-412-05551-1.
- [12] H. Haario, E. Saksman, and J. Tamminen. An adaptive Metropolis algorithm. *Bernoulli*, 7(2):223–242, 2001.
- [13] U. H. E. Hansmann. Parallel tempering algorithm for conformational studies of biological molecules. *Chem. Phys. Lett.*, 281(1-3):140–150, 1997.
- [14] D. M. Higdon. Auxiliary variable methods for Markov chain Monte Carlo with applications. *J. Amer. Statist. Assoc.*, 93:585–595, 1997.
- [15] S. F. Jarner and E. Hansen. Geometric ergodicity of Metropolis algorithms. *Stochastic Process. Appl.*, 85(2):341–361, 2000.
- [16] D. Kofke. On the acceptance probability of replica-exchange Monte Carlo trials. *J. Chem. Phys.*, 117(15):6911–1914, 2002.
- [17] A. Kone and D. Kofke. Selection of temperature intervals for parallel-tempering simulations. *J. Chem. Phys.*, 122, 2005.
- [18] H. Kushner and G. Yin. *Stochastic Approximation and Recursive Algorithms and Applications*. Springer, 2003.
- [19] F. Liang and W. H. Wong. Evolutionary Monte Carlo for protein folding simulations. *J. Chem. Phys.*, 115(7), 2001.
- [20] E. Marinari and G. Parisi. Simulated tempering: a new Monte Carlo scheme. *Europhys. Lett.*, 19(6):451–458, 1992.
- [21] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equations of state calculations by fast computing machines. *J. Chem. Phys.*, 21(6):1087–1092, 1953.



- [22] S. Meyn and R. L. Tweedie. *Markov Chains and Stochastic Stability*. Cambridge University Press, New York, NY, USA, 2nd edition, 2009. ISBN 0521731828, 9780521731829.
- [23] S. P. Meyn and R. L. Tweedie. Computable bounds for geometric convergence rates of Markov chains. *Ann. Appl. Probab.*, 4(4): 981–1011, 1994.
- [24] C. P. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer-Verlag, New York, 1999. ISBN 0-387-98707-X.
- [25] G. O. Roberts and J. S. Rosenthal. Optimal scaling for various Metropolis-Hastings algorithms. *Statist. Sci.*, 16(4):351–367, 2001.
- [26] G. O. Roberts and J. S. Rosenthal. Examples of adaptive MCMC. *J. Comput. Graph. Statist.*, 18(2):349–367, 2009.
- [27] G. O. Roberts and J. S. Rosenthal. Minimising MCMC variance via diffusion limits, with an application to simulated tempering. Technical report, 2012. URL <http://probability.ca/jeff/research.html>.
- [28] G. O. Roberts, A. Gelman, and W. R. Gilks. Weak convergence and optimal scaling of random walk Metropolis algorithms. *Ann. Appl. Probab.*, 7(1):110–120, 1997.
- [29] E. Saksman and M. Vihola. On the ergodicity of the adaptive Metropolis algorithm on unbounded domains. *Ann. Appl. Probab.*, 20(6):2178–2203, 2010.
- [30] R. H. Swendsen and J.-S. Wang. Replica Monte Carlo simulation of spin-glasses. *Phys. Rev. Lett.*, 57(21):2607–2609, 1986.
- [31] M. Vihola. On the stability and ergodicity of adaptive scaling Metropolis algorithms. *Stochastic Process. Appl.*, 121(12):2839–2860, 2011.
- [32] M. Vihola. Can the adaptive Metropolis algorithm collapse without the covariance lower bound? *Electron. J. Probab.*, 16:45–75, 2011.
- [33] M. Vihola. Robust adaptive Metropolis algorithm with coerced acceptance rate. *Statist. Comput.*, 2012. to appear.
- [34] D. B. Woodard, S.C. Schmidler, and M. Huber. Conditions for rapid mixing of parallel and simulated tempering on multimodal distributions. *Ann. Appl. Probab.*, 19(2):617–640, 2009.

BLAŻEJ MIASOJEDOW, LTCI, CNRS UMR 8151, INSTITUT MINES-TÉLÉCOM/TÉLÉCOM  
PARISTECH, 46, RUE BARRAULT, 75634 PARIS CEDEX 13, FRANCE

ERIC MOULINES, LTCI, CNRS UMR 8151, INSTITUT MINES-TÉLÉCOM/TÉLÉCOM  
PARISTECH, 46, RUE BARRAULT, 75634 PARIS CEDEX 13, FRANCE

MATTI VIHOLA, DEPARTMENT OF MATHEMATICS AND STATISTICS, P.O.Box  
35, FI-40014 UNIVERSITY OF JYVÄSKYLÄ, FINLAND